



UNIVERSAL  
LIBRARY



103 293

UNIVERSAL  
LIBRARY

**HOGBEN / STATISTICAL THEORY**

# **STATISTIC THEORY**

*A relationship between  
probability and statistics*

**LANCET  
HOGBEN, F.R.**

# Statistical Theory

THE RELATIONSHIP OF PROBABILITY,  
CREDIBILITY, AND ERROR

*By* LANCELOT HOGBEN, F.R.S.

Statistical theory is of importance to the plain man as well as to the trained mathematician. Scientists, sociologists, and administrators show an increasing disposition to exploit the newest statistical devices with little concern for their mathematical credentials or the formal assumptions contained within them.

Writing as a biologist and a behavioralist, Professor Hogben examines the underlying assumptions of a statistical theory. He sets down, with the clarity, brilliance, and force one expects from him, the views of the scientist who *uses* the formulae of statistics. He distinguishes four elements in today's theory of statistics: a calculus of error, a calculus of aggregates, a calculus of exploration, and a calculus of judgments, and he examines all of them according to their origins, contents, and validity. By tracing current conflicts of doctrine to their sources, he makes clear to the younger generation of research workers how important it is to examine and to question the credentials of principles invoked in the course of their work.



KANSAS CITY, MO. PUBLIC LIBRARY



3 1148 01171 0688

kansas city



public library

kansas city, missouri

Books will be issued only  
on presentation of library card.  
Please report lost cards and  
change of residence promptly.  
Card holders are responsible for  
all books, records, films, pictures  
or other library materials  
checked out on their cards.

# STATISTICAL THEORY

The Relationship of  
Probability, Credibility and Error

To  
R. F. W.

LANCELOT HOGBEN, F.R.S.

# *Statistical Theory*

The Relationship of  
PROBABILITY, CREDIBILITY  
AND ERROR

AN EXAMINATION OF THE CONTEMPORARY  
CRISIS IN STATISTICAL THEORY  
FROM A BEHAVIOURIST VIEWPOINT

---

W · W · NORTON & COMPANY INC  
Publishers New York

**All rights reserved.**  
**Published simultaneously in Canada by**  
**George J. McLeod Limited, Toronto.**

**Printed in the United States of America**

**1 2 3 4 5 6 7 8 9 0**

# CONTENTS

Foreword	page 7
1 <i>The Contemporary Crisis or the Uncertainties of Uncertain Inference</i>	13
PART I. THE FOUNDING FATHERS	
2 <i>The Founding Fathers and the Natural History of Gambling</i>	33
3 <i>Randomness and the Relevance of the Rule</i>	59
4 <i>Division of the Stakes and the Lottery of Life and Death</i>	83
5 <i>The Backward Look and the Balance Sheet of Thomas Bayes</i>	110
6 <i>The Place of Laplace and the Grand Climacteric of the Classical Tradition</i>	133
PART II. THE CALCULUS OF ERROR AND THE CALCULUS OF EXPLORATION	
7 <i>The Normal Law comes into the Picture</i>	159
8 <i>The Method of Least Squares and the Concept of Point Estimation</i>	182
9 <i>Error, Variation and Natural Law</i>	210
10 <i>Stochastic Models for Simple Regression</i>	232
11 <i>The Impasse of Factor Analysis</i>	257
PART III. THE CALCULUS OF AGGREGATES	
12 <i>Maxwell and the Urn of Nature</i>	279
13 <i>Mendelism and the Two Meanings of Probability</i>	297
PART IV. THE CALCULUS OF JUDGMENTS	
14 <i>Statistical Prudence and Statistical Inference</i>	319
15 <i>Decision, Indecision and Sample Economy</i>	345
16 <i>Induction and Design, Stochastic and Non-stochastic</i>	370
17 <i>Recipe and Rule in Stochastic Induction</i>	399
18 <i>The Confidence Controversy and the Fifth Canon</i>	433
19 <i>Epilogue</i>	454
Appendix I	477
Appendix II	480
Appendix III	485
Appendix IV	487
Index	07





## FOREWORD

THE USE of the word behaviourist in my sub-title calls for clarification. It came into current usage chiefly through the writings of J. B. Watson, in the first fine flush of enthusiasm following the reception of Pavlov's work on conditioned reflexes. Watson conveyed the impression that all forms of animal—including human—behaviour are ultimately explicable in terms of neural or humoral co-ordination of receptors and effectors. Neither this proposition nor its denial is susceptible of proof. A comet may destroy the earth before we have completed a research programme of such magnitude as the operative adverb *ultimately* suggests.

Many years ago, and in what now seems to me a very immature volume of essays\* written as a counter-irritant to the mystique of Eddington's *Nature of the Physical World*, I suggested a more restricted definition of the term and offered an alternative which seemingly did not commend itself to my contemporaries. By *behaviourist* in this context I mean what Ryle means in the concluding section of his recent book *The Concept of Mind*. What Ryle speaks of as the behaviourist viewpoint and what I had previously preferred to call the *publicist* outlook has no concern with structure and mechanism. Our common approach to the problem of cognition is not at the ontological level. The class of questions we regard as profitable topics of enquiry in this context arise at the level of epistemology, or, better still, what G. P. Meredith calls *epistemics*. If one wishes to discuss with any hope of reaching agreement what one means by knowledge—or preferably *knowing*—the two main topics of the agenda for our *public* symposium will be *recognition* and *communication*; and we shall discuss them as different facets of the same problem, neither meaningful in isolation. A simple illustration will suffice to make clear, both at the emotive level of non-communicable private conviction and at the public or communicable level of observable behaviour, the difference between *knowing*, conceived as a process, and *knowledge*, conceived as a static repository.

\* *The Nature of Living Matter* (1930).

We shall suppose that B is colour-blind to differences in the red and green regions of the visible spectrum, A being normal in the customary sense of the term in the relevant context. Faced with a crimson Camellia and a blade of grass B asserts that they differ only with respect to light and shade, shape and texture, possibly or probably also with respect to chemical composition, microscopic structure and other characteristics less accessible to immediate inspection. A admits everything in this assertion except the implications of the word *only*. He *knows* that red is red and green is green, just as surely as B *knows* that the distinction is frivolous. The deadlock is complete till they call in C who suggests that it is possible to submit the subject-matter of their disagreement to the outcome of an identity parade. At this point, we shall assume that A and B, each equally secure in his convictions, both agree to do so. The proposal C offers takes shape as follows. First, each disputant inspects without handling a set of objects of like shape arranged in a row. Some are green and some are red. A labels the red ones  $R_1, R_2, R_3 \dots$  etc., and the green  $G_1, G_2, G_3 \dots$  etc. B denies that they are distinguishable. The umpire C now photographs the set twice in the presence of A and B, using first an Ilford filter which cuts out all rays in the green region and then one which cuts out all rays at the red end of the spectrum. He invites both A and B to watch while he develops and fixes the film. The reader may complete the rest of the parable. All that remains is to point out the moral.

If we are content to discuss the contemporary status of the calculus of probability as an instrument of research in terms of what goes on *in the mind*, we have as little hope of arriving at agreement as had A and B before the advent of C. If we seek to find a highest common factor for the process of knowing, we must therefore find some means of submitting our differences to an identity parade. We shall then cease to chatter in the private idiom of conviction and begin to communicate in the public vocabulary of events. Needless to say, we cannot force either B or A to do this. We can merely invite them to do so. As Ryle would put it, the utmost we can do in the context of the topic of the pages which follow is to make *linguistic recommendations* which will help them to resolve their differences, if willing

## FOREWORD

to do so. At the outset, we must therefore insist on some acceptable distinction between *knowing* and *believing*.

The behaviourist, as I here use the term, does not deny the convenience of classifying *processes* as mental or material. He recognises the distinction between personality and corpse; but he has not yet had the privilege of attending an identity parade in which human minds without bodies are by common recognition distinguishable from living human bodies without minds. Till then, he is content to discuss probability in the vocabulary of *events*, including audible or visibly recorded assertions of human beings as such. He can concede no profit from a protracted debate at the level of what goes on *in* the mind. The use of the italicised preposition in the idiom of his opponents is indeed inconsistent with any assertion translatable in terms of a distinction C can induce A and B to inspect; and this dichotomy is not one which concerns the mathematician alone. It is not even an issue on which prolonged preoccupation with intricate symbolic operations divorced from contact with living things necessarily equips a person to arbitrate with a compelling title to a respectful hearing. None the less, it is one which forces itself on our attention from whatever aspect we view the contemporary crisis in Statistical Theory.

Hitherto, the contestants have been trained mathematicians, many of whom explicitly, and many more of whom implicitly, reject the behaviourist approach outlined in the preceding paragraphs. The writer is a trained biologist. Like most other trained biologists, he has a stake in knowing how far, and for what reasons, statistical theory can help us to enlarge our understanding of living matter. Since mathematical statisticians have not been hesitant to advance ambitious claims about the help it can give them, it is not presumptuous for a biologist to examine the credentials of such claims in the language of behaviour, the only language which all biologists share. If statisticians wish to decline, as Carnap most explicitly does decline, to communicate with the biologist on this level of discourse, they are still free to do so on this side of the Iron Curtain; and, happily, there are publishers who will make their views accessible to the consumer. It is entirely defensible to formulate an axiomatic approach to

the theory of probability as an internally consistent set of propositions, if one is content to leave to those in closer contact with external reality the last word on the usefulness of the outcome.

What invests the present controversies concerning the foundations of theoretical statistics with a peculiar flavour of comedy is a hyper-sensitive anxiety to publicise the relevance of the outcome to the world of affairs on the part of those who repudiate most vigorously the relevance of the external world to the choice of their axioms; but it is by no means true to represent the current crisis in statistical theory as a contest between opposing camps in one of which all the contestants are Platonists, in the other all of them behaviourists as defined above. The truth is otherwise. In one sense, we are all behaviourists nowadays. At least, few persons outside a mental home would defend thoroughgoing Platonic idealism. What is equally true is that all of us carry over from the past some remnants of mental habits which antedate the naturalistic outlook of our times. Hence we may consistently assert the relevance of the dichotomy to the content of the contemporary controversy while expecting to encounter among the disputants no greater measure of consistent adherence to one or other viewpoint than experience of the inconsistencies of gifted human beings can justify. Even Carnap, whom I have cited as a protagonist of the axiomatic school, does so with a generous concession to his own day and age by distinguishing between probability 1 and probability 2, or in my own idiom private and public probability.

This book is a prosaic discussion about public probability. If the topic were private, the author would have communicated his convictions through the medium of verse composition. One concern of the writer in the course of revaluating his own views by tracing current conflicts of doctrine to their sources has been to remedy an increasingly widespread disposition of the younger generation of research workers to relinquish the obligation to examine the credentials of principles invoked in the day's work. Indeed, one may well ask whether a liberal education in theology is nowadays less conducive to acquiescence in authoritarian dogma than is rule of thumb instruction in statistical methods as now commonly given to students of

## FOREWORD

biology and the social sciences. Such acquiescence is no doubt due in part to the formidable algebra invoked by the theoreticians of statistics during the inter-war decades. For that reason, I have attempted in an earlier book to exploit a novel technique of visual aids to make accessible to a wider audience than those to which professional mathematicians commonly address themselves, the formal algebra of sampling distributions frequently referred to in this book. I trust that *Chance and Choice* will still serve some useful purpose on that account; but common honesty compels me to repudiate its claim to have accomplished an overdue revaluation of statistical theory.

In carrying out the self-imposed task attempted therein, my own views about widely accepted teaching became more and more critical as examination of current procedures against a background of explicitly visualisable stochastic models disclosed factual assumptions too easily concealed behind an impressive façade of algebraic symbols; but my concluding remarks were tentative and will have left the reader with what I now regard as an unduly optimistic anticipation of legitimate compromise. Here my aim has been to take stock of the outcome and to make explicit a viewpoint much less conservative than that of *Chance and Choice*. My present views are in direct contrariety to many therein expressed in conformity with current tradition.

Thus this book is not a textbook, complete in itself as a course of study. On the other hand, avoidance of any exposition of relevant algebraic expressions would have been inconsistent with the intention stated. To be sure, the lay-out would have been more pleasing, if it had been advisable to allocate more equably space devoted to historical and methodological commentary on the one hand and to algebraic derivations on the other. The writer has resisted the temptation to sacrifice the dictates of clarity to the aesthetics of book-making, because it has seemed best to steer a middle course between encouraging the reader to take too much on trust and recapitulating demonstrations to which he or she has now ready access. Thus the reader who commendably declines to take the algebraic theory of the  $t$ -test on trust may refer to Kendall's *Advanced Theory* or, if unable to follow all the steps set out therein, may consult Chapters 15 and 16 of Vol. II of *Chance and Choice*.

Contrariwise, Chapters 6–9 of this book have a more textbook flavour than most of those which precede or follow. I have written them in this way because biologists and sociologists will find little if any reference to the Gaussian Theory of Error in statistical textbooks addressed to themselves. Accordingly, they will not readily retrace the ancestry of current procedures subsumed under the terms regression and multivariate analysis unless forearmed by an explicit exposition of the method of least squares in its first domain of application.

## CHAPTER ONE

# THE CONTEMPORARY CRISIS OR THE UNCERTAINTIES OF UNCERTAIN INFERENCE

IT IS NOT without reason that the professional philosopher and the plain man can now make common cause in a suspicious attitude towards statistics, a term which has at least five radically different meanings in common usage, and at least four in the context of *statistical theory* alone. We witness on every side a feverish concern of biologists, sociologists and civil servants to exploit the newest and most sophisticated statistical devices with little concern for their mathematical credentials or for the formal assumptions inherent therein. We are some of us all too tired of hearing from the pundits of popular science that natural knowledge has repudiated any aspirations to absolute truth and now recognises no universal logic other than the principles of statistics. The assertion is manifestly false unless we deprive all purely taxonomic enquiry of the title to rank as science. It is also misleading because statistics, as men of science use the term, may mean disciplines with little connexion other than reliance, for very different ostensible reasons, on the same algebraic tricks.

This state of affairs would be more alarming as indicative of the capitulation of the scientific spirit to the authoritarian temper of our time, if it were easy to assemble in one room three theoretical statisticians who agree about the fundamentals of their speciality at the most elementary level. After a generation of prodigious proliferation of statistical techniques whose derivation is a closed book to an ever-expanding company of avid consumers without access to any sufficiently simple exposition of their implications to the producer-mathematician, the challenge of J. Neyman, E. S. Pearson and Abraham Wald is provoking, in Nietzsche's phrase, a transvaluation of all values. Indeed, it is not too much to say that it threatens to undermine the entire superstructure of statistical estimation and test procedure erected by R. A. Fisher and his disciples on

the foundations laid by Karl Pearson, Edgeworth and Udny Yule. An immediate and hopeful consequence of the fact that the disputants disagree about the factual credentials of even the mathematical theory of probability itself is that there is now a market for textbooks on probability as such, an overdue awareness of its belated intrusion in the domain of scientific research and a willingness to re-examine the preoccupations of the Founding Fathers when the topic had as yet no practical interest other than the gains and losses of a dissolute nobility at the gaming table.

Since unduly pretentious claims put forward for statistics as a discipline derive a spurious cogency from the protean implications of the word itself, let us here take a look at the several meanings it enjoys in current usage. First, we may speak of statistics in a sense which tallies most closely with its original connotation, i.e. figures pertaining to affairs of state. Such are *factual* statistics, i.e. any body of data collected with a view to reaching conclusions referable to recorded numbers or measurements. We sometimes use the term *vital* statistics in this sense, but for a more restricted class of data, e.g. births, deaths, marriages, sickness, accidents and other happenings common to individual human beings and more or less relevant to medicine, in contradistinction to information about trade, employment, education and other topics allocated to the social sciences. In a more restricted sense, we also use expressions such as vital statistics or economic statistics for the exposition of *summarising* procedures (life expectation, age standardisation, gross or net reproduction rates, cohort analysis, cost of living or price indices) especially relevant to the analysis of data so described. By analysis in this context, we then mean sifting by recourse to common sense and simple arithmetical procedures what facts are more or less relevant to conclusions we seek to draw, and what circumstances might distort the true picture of the situation. Anscombe (*Mind*, 1951) refers to analysis of this sort as statistics in the sense in which "some continental demographers" use the term.

If we emphatically repudiate the unprovoked scorn in the remark last cited, we must agree with Anscombe in one particular. When we speak of analysis in the context of demography,



we do not mean what we now commonly call *theoretical statistics*. What we do subsume under the latter presupposes that our analysis invokes the calculus of probabilities. When we speak of a calculus of probabilities we also presuppose a single formal system of algebra; but a little reflection upon the history of the subject suffices to remind us that: (a) there has been much disagreement about the relevance of such a calculus to everyday life; (b) scientific workers invoke it in domains of discourse which have no very obvious connexion. When I say this I wish to make it clear that I do not exclude the possibility that we may be able to clarify a connexion if such exists, but only if we can reach some agreement about the relevance of the common calculus to the world of experience. On that understanding, we may provisionally distinguish between four domains to which we may refer when we speak of the *Theory of Statistics*:

(i) *A Calculus of Errors*, as first propounded by Legendre, Laplace and Gauss, undertakes to prescribe a way of combining observations to derive a preferred and so-called *best* approximation to an assumed *true* value of a dimension or constant embodied in an independently established law of nature. The algebraic theory of probability intrudes at two levels: (a) the attempt to interpret empirical laws of error distribution referable to a long sequence of observations in terms consistent with the properties of models suggested by games of chance; (b) the less debatable proposition that unavoidable observed net error associated with an isolated observation is itself a sample of elementary components selected randomwise in accordance with the assumed properties of such models.

Few current treatises on theoretical statistics have much to say about the Gaussian Theory of Error; and the reader in search of an authoritative exposition must needs consult standard texts on *The Combination of Observations* addressed in the main to students of astronomy and geodesics. In view of assertions mentioned in the opening paragraph of this chapter, it is pertinent to remark that a calculus for combining observations as propounded by Laplace and by Gauss, and as interpreted by all their successors, presupposes a putative true value of any measurement or constant under discussion as a secure foothold

for the concept of error. When expositors of the contemporary reorientation of physical theory equate the assertion that the canonical form of the scientific law is statistical to the assertion that the new physicist repudiates absolute truth, cause and effect as irrelevant assumptions, it is therefore evident that they do not use the term statistical to cover the earliest extensive application of the theory of probability in the experimental sciences. I cannot therefore share with my friend Dr. Bronowski the conviction that the statistical formulation of particular scientific hypotheses subsumed under the calculus of aggregates as defined below has emancipated science from an aspiration so old-fashioned as the pursuit of absolute truth. Still less do I derive from so widely current a delusion any satisfaction from the promise of a new Elizabethan era with invigorating prospects of unforeseen mental adventure.

(ii) *A Calculus of Aggregates* proceeds deductively from certain axioms about the random behaviour of subsensory particles to the derivation of general principles which stand or fall by their adequacy to describe the behaviour of matter in bulk. In this context our criterion of adequacy is the standard of precision we commonly adopt in conformity with operational requirements in the chosen field of enquiry. Clerk Maxwell's Kinetic Theory of Gases is the *fons et origo* of this prescription for the construction of a scientific hypothesis and the parent of what we now call statistical mechanics. Beside it, we may also place the Mendelian Theory of Populations.

So far as I know, the reader in search of an adequate account of statistical hypothesis of this type will not be able to find one in any standard current treatise ostensibly devoted to Statistical Theory as a whole. This omission is defensible in so far as physicists and biologists admittedly accept the credentials of such hypotheses on the same terms as they accept hypotheses which make no contact with the concept of probability, e.g. the thermodynamic theory of the Donnan membrane equilibrium. We assent to them because they *work*. Seemingly, it is unnecessary to say more than that, since all scientific workers agree about what they mean in the laboratory, when they say that a hypothesis works; but such unanimity has no bearing on the plea that statistical theory works, when statistical

theory signifies the contents of contemporary manuals setting forth a regimen of interpretation now deemed to be indispensable to the conduct of research in the biological and social sciences.

(iii) *A Calculus of Exploration*, which I here use to label such themes as regression and factor analysis, is difficult to define without endorsing or disclaiming its credentials. The expression is appropriate in so far as the ostensible aim of procedures subsumed as such is definable in the idiom of Karl Pearson as *concise statement of unsuspected regularities of nature*. This again is vague, but less so if we interpret it in terms of Pearson's intention, and of Quetelet's belief that social phenomena are subject to quantitative laws as inexorable as those of Kepler. Procedures we may designate in this way, more especially the analysis of covariance, may invoke significance tests, and therefore intrude into the domain of the calculus of judgments; but the level at which the theory of probability is ostensibly relevant to the original terms of reference of such exploratory techniques is an issue *sui generis*.

The pivotal concept in the algebraic theory of regression, as in the Gaussian theory of error, is that of *point-estimation*; and the two theories are indeed formally identical. In what circumstances this extension of the original terms of reference of the calculus of errors took place is worthy of comment at an early stage. We shall later see how a highly debatable transference of the theory of wagers in games of chance to the uses of the life table for assessment of insurance risks whetted the appetite for novel applications of the algebraic theory of probability in the half-century before the more mature publications of Gauss appeared in print. The announcement of the Gaussian theory itself coincided with the creation of new public instruments for collection of demographic data relevant to actuarial practice both in Britain\* and on the Continent. Therewith we witness the emergence of the professional statistician in search of a theory. Such was the setting in which Quetelet obtained a considerable following, despite the derision of Bertrand (p. 172)

\* The Office of the Registrar-General of England and Wales came into being in 1837.

and other mathematicians *au fait* with the factual assumptions on which the Gaussian theory relies.

Since an onslaught by Keynes, the name of Quetelet, so explicitly and repeatedly acknowledged by Galton, by Pearson and by Edgeworth as the parent of regression and cognate statistical devices, has unobtrusively retreated from the pages of statistical textbooks; but mathematicians of an earlier vintage than Pearson or Edgeworth had no illusions about the source of his theoretical claims nor about the relevance of the principles he invoked to the end in view. So discreet a disinclination to probe into the beginnings of much we now encounter in an up-to-date treatise on statistical theory will not necessarily puzzle the reader, if sufficiently acquainted with the unresolved difficulties of reaching unanimity with respect to the credentials of the remaining topics there dealt with; but we shall fail to do justice to the legitimate claims of its contents, unless we first get a backstage view of otherwise concealed assumptions. In what seems to be one of the first manuals setting forth the significance test drill, Caradog Jones (*First Course in Statistics*, 1921) correctly expounds as follows the teaching of Quetelet and its genesis as the source of the tradition successively transmitted through Galton, Pearson and R. A. Fisher to our own contemporaries:

It is almost true to say, however, that until the time of the great Belgian, Quetelet (1796–1874), no substantial theory of statistics existed. The justice of this claim will be recognized when we remark that it was he who really grasped the significance of one of the fundamental principles—sometimes spoken of as the *constancy of great numbers*—upon which the theory is based. A simple illustration will explain the nature of this important idea: imagine 100,000 Englishmen, all of the same age and living under the same normal conditions—ruling out, that is, such abnormalities as are occasioned by wars, famines, pestilence, etc. Let us divide these men at random into ten groups, containing 10,000 each, and note the age of every man when he dies. Quetelet's principle lays down that, although we cannot foretell how long any particular individual will live, the ages at death of the 10,000 added together, whichever group we consider, will be practically the same. Depending upon this fact, insurance companies calculate the premiums they must charge, by a process of averaging mortality results recorded in the past, and so they are able to carry on business without serious risk

of bankruptcy. . . . In his writings he visualizes a man with qualities of average measurement, physical and mental (*l'homme moyen*), and shows how all other men, in respect of any particular organ or character, can be ranged about the mean of all the observations. Hence he concluded that the methods of Probability, which are so effective in discussing errors of observation, could be used also in Statistics, and that deviations from the mean in both cases would be subject to the binomial law.

If we are to do justice to the claims of a calculus of exploration, we must therefore ask in what sense probability is indeed so effective in discussing errors of observation and in what sense, if any, are Quetelet's authentic deviations from a non-existent population mean comparable to the Gaussian deviation of a measurement from its putatively authentic value. We shall then envisage the present crisis in statistical theory as an invitation to a more exacting re-examination of its foundations than contemporary controversy has hitherto encompassed. After the appearance of his treatise on probability by Keynes, who dismisses Quetelet as a charlatan with less than charity towards so many highly esteemed contemporaries and successors seduced by his teaching, an open conspiracy of silence has seemingly exempted a younger generation from familiarity with the thought of the most influential of all writers on the claims of statistical theory in the world of affairs. Since his views will occupy our attention again and again in what follows, a few remarks upon his career, culled from Joseph Lottin's biography and from other sources, will be appropriate at the outset.

From 1820 onwards Quetelet was director of the Royal Belgian Observatory which he founded. In the 'twenties he professed astronomy and geodesy at the *Ecole Militaire*. The year following the publication (1835) of his portentous *Essai de Physique Sociale*, their uncle King Leopold committed to his care Albert of Saxe-Coburg and his brother Ernest for a brief course of instruction on the principles of probability. Correspondence continued between Quetelet and the Prince, who remained his enthusiastic disciple, affirming his devotion to the doctrine of the *Essai* both as president of the 1859 meeting of the British Association in Aberdeen when Maxwell first announced his stochastic interpretation of the gas laws and in the next year

as president of the International Statistical Congress held (1860) in London. As official Belgian delegate, Quetelet himself had attended (1832) the third annual meeting of the British Association at Cambridge. There he conferred with Malthus and Babbage, then Lucasian professor of the Newtonian succession and the inventor of the first automatic computer, also famous as author of the *Economy of Manufacture* and of the *Decline of Science in England*. The outcome of their deliberations was the decision of the General Committee to set up a statistical section.

An incident in the course of Quetelet's relations with Albert is revealing and not without entertainment value. Gossart (1919) tells of it thus in the *Bulletin de la Classe des Lettres*, etc., of the Royal Belgian Academy.

Quetelet peu après la publication de ses lettres; presenta à l'Academie un mémoire *Sur la Statistique morale et les principes qui doivent en former la base*. . . . Par une fâcheuse coïncidence, le volume dans lequel les théories de Quetelet touchaient à l'art de gouvernement paraissait à Paris au moment on éclatait la révolution de février 1848 et allait "se perdre au milieu des barricades" si bien que quelques exemplaires seulement furent alors distribués. En voyant ce que se passait en France et bientôt dans une partie de l'Europe, le prince Albert ne put s'empêcher de remarquer avec une certaine pointe de malice que le système social était "bien dérangé," que les "causes accidentelles" jouaient un grand rôle. "Le malheur," ajoutait il, "est que la loi qui les gouverne n'a pas été decouverte jusqu'à ce moment."

Quetelet's belief in eternal laws of human society was *en rapport* with a social philosophy unruffled by such mishaps as the Commune; and its Calvinistic temper is hard to reconcile with the libertarian *credo* for which Eddington finds sanction by appeal to the principle of uncertainty lately propounded by the exponents of statistical mechanics. "Tout en déplorant les maux que font à la société 'les changements brusque et les théories des rêveurs'," says Gossart, he attained *la resignation*. To be sure, "des fleaux frappent l'humanité au morale comme au physique," as he admits; but "quelque destructifs que soient leurs effets, il est au moins consolant de penser qu'ils ne peuvent altérer en rien les lois éternelles qui nous régissent. Leur action est

passagere et le temps a bientôt cicatrisé les plaies du corps social."

(iv) *A Calculus of Judgments*, as here defined, ostensibly embraces a regimen of correct inference with respect to the credentials of hypotheses. This form of words is advisedly vague, because it is impossible to prescribe its terms of reference more explicitly without prejudging the outcome of the contemporary controversy we are about to examine. If one states that it includes both the theory of significance and of decision tests and the theory of interval estimation in terms of confidence or fiducial limits, the reader will infer all that we need say definitively and with propriety at this stage.

As such, the calculus of judgments subsumes a programme which is almost exclusively a product of our own century; but the emergence of the aspiration the programme endorses is traceable to the doctrine of inverse probability adumbrated in the posthumous publication of Bayes (1763) and propounded more explicitly by Laplace. The end Laplace himself had in view was to vindicate the credentials of inductive reasoning conceived in retrospective terms consistent with his own cosmogony and hence likewise in terms of dubious relevance to verification in the domain of experimental design. Most statisticians now reject the doctrine in its original form; but its essentially retrospective orientation is profoundly relevant to the contemporary crisis in statistical theory.

At what level the theory of probability can appropriately intrude in a prescription for reasoning rightly is an issue which we can discuss with profit if, and only if, we can arrive at agreement about the relevance of the theory of probability to induction in the traditional sense of the term. This is not the exclusive prerogative of the mathematician. It is the birthright and duty of every self-respecting scientific worker who subjects his data to the type of analysis prescribed by one or other school of statistical inference. That fundamental differences with respect to the relevance of the mathematical theory to the world of real decisions have come into focus so lately is not surprising, when we reflect on the circumstances that its application to the technique of interpretation basks in the reflected glory of the pragmatic triumphs of Maxwell, Mendel and their successors.

In the deductive unfolding of a theory which must stand or fall with the operational requirements of laboratory practice, we are entitled to start from any axioms however arbitrary. We should therefore scrutinise with some suspicion the following remarks about statistical methods by Wilks (1944):

The test of the applicability of the mathematics in this field as in any other branch of mathematics consists in comparing the predictions as calculated from the mathematical model with what actually happens experimentally. (*Mathematical Statistics*, Chapter 1, p. 1.)

This assertion would be unexceptionable, if statisticians invoked the algebra of Professor Wilks only in connexion with the genetical theory of populations, Brownian movement of visible particles, the collisions of gas molecules, the emission of photons, and with cognate themes which constitute the scope of a calculus of aggregates; but such topics have no direct relevance to what we imply by statistical theory in the context of a calculus of judgments. In the calculus of aggregates we invoke the theory of probability to prescribe a hypothesis; but a calculus of judgments does not undertake to prescribe hypotheses. It claims only to prescribe a rule which will entitle us to arbitrate on their merits. One may hence ask with propriety what controlled experiment prosecuted on either side of what iron curtain over what number of centuries would settle the dispute between Jeffreys and Fisher concerning the legitimacy of Bayes's postulate or the contest between Fisher and Neyman over test procedure. When the terrain of combat is the realm of means, experience and experience alone should dictate the outcome. When it is in the realm of ends we cannot invoke pragmatic sanctions with the assurance of an acceptable decision.

\* \* \*

In the chapters which follow it will be the writer's aim to set forth the terms of reference of the classical theory in some detail at the outset and thereafter to examine its bearing on the four main themes distinguished in the foregoing paragraphs. First, I shall invite the reader to agree with me that my sub-title does not overstate what is a real intellectual dilemma of our time.



*Crisis* is a word which has become tarnished by misuse; and some of my readers may well wish me to justify the statement that there is indeed a contemporary crisis in statistical theory. Poincaré cites a remark that everyone believes in the normal law of error, the physicists because they think that the mathematicians have proved it to be a logical necessity, the mathematicians because they believe that physicists have established it by laboratory demonstration. The gap between theory and practice has vastly deepened since his time, as is evident from the concluding remarks of E. S. Pearson (1944)\* in the following excerpt:

That the frequency concept is not generally accepted in the interpretation of statistical tests is of course well known. With his characteristic forcefulness R. A. Fisher (1945b) has recently written: "In recent times one often repeated exposition of the tests of significance, by J. Neyman, a writer not closely associated with the development of these tests, seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom, which is foreign to the reasoning on which the tests of significance were in fact based, seems to be a real bar to progress. . . ."

But the subject of criticism seems to me less an intrusive mathematical axiom than a mathematical formulation of a practical requirement which statisticians of many schools of thought have deliberately advanced. Prof. Fisher's contributions to the development of tests of significance have been outstanding, but such tests, if under another name, were discovered before his day and are being derived far and wide to meet new needs. To claim what seems to amount to patent rights over their interpretation can hardly be his serious intention. Many of us, as statisticians, fall into the all too easy habit of making authoritative statements as to how probability theory should be used as a guide to judgment, but ultimately it is likely that the method of application which finds greatest favour will be that which through its simplicity and directness appeals most to the common scientific user's understanding. *Hitherto the user has been accustomed to accept the function of probability theory laid down by the mathematicians; but it would be good if he could take a larger*

\* *Biometrika*, Vol. 34, p. 142.

*share in formulating himself what are the practical requirements that the theory should satisfy in application."* (Italics inserted.)

Meanwhile the user, as Pearson calls him, continues to perform an elaborate ritual of calculations quite regardless of the fact that there are now at least three schools of theoretical doctrine with no common ground concerning what justification we have for applying a calculus of probability to real situations and with little agreement about how we should proceed to do so. Lest some readers should regard this as an overstatement, it will not be amiss to quote from a recent symposium following a paper read before the Royal Statistical Society\* by Anscombe (1948) who undertook the courageous assignment of an impartial appraisal of the views respectively advanced during the last decade by R. A. Fisher, by H. Jeffreys and by J. Neyman and E. S. Pearson. Dr. J. O. Irwin (p. 201) who opened the discussion said:

I think all students of statistics should learn something about probability from a frequency point of view. When teaching students with mature minds who are yet new or almost new to the subject, I usually give an outline of the different theories of the subject, tell them that they will find the frequency theory the most useful in practice, and to suspend judgment on which theory they will ultimately prefer as a basis until they have had more opportunity of study.

Practically minded people with no great taste for logical and philosophical speculation need not probe too deeply. They will probably be just as good statisticians if they don't. More theoretically profound minds will gain much insight if they do and will be able to help the others on critical occasions. But we must admit that what level we agree to call axiomatic is largely a matter of taste.

Professor G. A. Barnard, who followed Dr. Irwin, said in a more explicitly accommodating vein (pp. 201-2):

All three main theories of statistical inference seem to have their proper sphere of application. What we should ask is, not so much which is right, but to what sort of field each theory should be applied; which framework is better in certain circumstances. For

\* Discussion of Mr. Anscombe's Paper, *J. Roy. Stat. Soc. (Series A)*, 1948, Vol. CXI.

example, in considering industrial inspection Dodge and Romig, two engineers, evolved the notions of producer's risk and consumer's risk, and these have been practically useful in statistical inspection. They are identical in content with the Pearson notion of errors of first and second kind. Again, during the war we had occasion to deal with other sorts of inspection problems, and in this connection we introduced the idea of the process curve, which is nothing but the *a priori* distribution of Professor Jeffreys. I must admit that in my own experience so far, cases where Professor Fisher's theory would have been most suitable have not been very frequent. I think that is because most of my problems have been those where it is necessary to make an administrative decision, rather than those in which one is concerned to establish or disprove a scientific theory. Our work is more concerned with immediate practical decisions, but I do not doubt that with wider experience I should have been able to quote practical cases for that also.

We are, then, left with three theories—the Jeffreys theory, the Pearson theory, and the Fisher theory. I think when we are discussing the foundations of statistics we should draw attention to the fact that this discussion is really, from a practical point of view, a discussion of the fine points of detail. All statisticians agree about what should be done in practical problems. The situation in statistics is really quite like the problem in mathematics. The foundations of mathematics have been discussed and queried for a long time. These discussions are now so broad and widespread that there is a journal devoted to them entirely. Yet no mathematician doubts that any of the mathematics are sound. . . . We have also to remember that a significance test, interpreted somewhat narrowly, as it must be, only allows us to say what is not true, but that does not involve proving a general proposition. In this connection a remark of Professor Jeffreys is worth quoting—that the methods of significance tests used in this way seem to enable us to disprove a great deal but never to prove anything. . . . I should like finally to make it clear that I disagree with all four parties to the controversy. The snag in Professor Jeffreys' theory is that to work it one has to specify a probability distribution for a class of alternative hypotheses and the whole of the probability has to be distributed. One must when interpreting one's experiments be able to think of all possible explanations of the data, and that, I think, none of us believe that we can do. It is always possible for someone to produce later an entirely new explanation we had never thought of, and which would not be represented in the hypothesis nor in the alternatives we had tested.

Taking that criterion it does suggest that in talking about inference from the probability point of view, leaving aside the rigorous ground of the randomization test, we ought to formulate our ideas, not in terms of probability, but in terms of odds. Bayes's theorem, in terms of odds, says that *a posteriori* odds = likelihood ratio (A)  $\times$  *a priori* odds. We can separate out the two factors on the right-hand side, and it seems to me that along these lines it is possible to reconcile to some extent the various theories. Professor Jeffreys takes the second factor as equal to one by a special axiom or assumption. Professor Fisher seems to say we ought to neglect the second factor; but that is equivalent to saying that A times 1 is equal to A. Finally, Neyman and Pearson say that A itself is a frequency probability of errors, and this is so provided that the reference set used is that of the sequential probability ratio test.

Reported in *oratio obliqua*, Professor E. S. Pearson (pp. 203-4) said with more insight into the consumer's viewpoint:

It has yet to be shown that a mathematical theory could make possible their assimilation into the process of inference on a numerical basis. In balancing these elements to reach conclusions leading to action the power of judgment was called into play; it was something which might be intuitive, a quality which scientific training aimed at developing, but whose possession was no monopoly of the statistician. The judgment might be an individual one or it might be a collective one according to the magnitude of the problem. The question which he raised was whether it was possible to lay down usefully any formal rules of induction, specifying how the various aspects of the problem needing review could be brought together to reach a balanced decision.

None of the contributors to this symposium advanced the most usual excuse for renouncing the traditional obligation of the man of science to understand what he is doing, i.e. the assertion that the procedures described by statistical theories work [*sic*] in spite of the fact that there is so exiguous a basis of agreement about their credentials. It is gratifying to record doubts about its cogency expressed by one of them in another context, if only because the consumer with no appetite for methodological disputation all too readily succumbs to reassurance on such terms. Acceptability of a statistically *significant* result of an experiment on animal behaviour in contra-

distinction to a result which the investigator can repeat before a critical audience naturally promotes a high output of publication. Hence the argument that the techniques *work* has a tempting appeal to young biologists, if harassed by their seniors to produce results, or if admonished by editors to conform to a prescribed ritual of analysis before publication. A reminder that the plea for justification by works derives its sanction from a different domain of statistical theory is therefore likely to fall on deaf ears, unless we reinstate reflective thinking in the university curriculum. Meanwhile, the views of E. S. Pearson\* on the teaching of statistics will commend themselves to the reflective few who entertain a pardonable scepticism about the allegedly useful contribution of current theories in so-called operational research:

Probably there are several of us who can recall a considerable number of reports, or appendices to reports, written on both sides of the Atlantic by mathematically trained statisticians, which were hardly more than a waste of the paper on which they were written. There were cases where the results of statistical analyses were simply put on one side because the practical man, whether scientist or service technician, shrewdly sensed that the theoretical treatment was either not needed, or was actually leading to conclusions which the data could not possibly warrant. The trouble usually arose because the mathematical enthusiast had allowed his theory to run away with his common sense, or, perhaps, because he had never received an adequate training in the application of theory. It was true that biologists did extremely well in operational research; but their success often seemed due to the way in which an experimental training had taught them to handle data rather than to the fact that they mastered statistical technique quickly.

I hope that these citations will dispel any doubt about whether there are very fundamental differences within the hierarchy of theoretical statistics concerning what the theory of probability can contribute to a regimen of scientific inference. They also disclose a widespread disposition on the part of the makers of the theory to disclaim at all costs any relevance of their differences to the requirements of those who use it.

\* Discussion on Dr. Wishart's Paper (*The Teaching of Statistics*), *J. Roy. Stat. Soc.*, 1948, Vol. CXI, p. 218.

Kendall (1949), who is deeply disturbed by the clamour the contemporary controversy has assumed, attempts in a recent paper "On the reconciliation of theories of Probability" to resolve disagreement by making explicit on what axioms widely current techniques in use depend; but his approach is essentially that of the pure mathematician seeking to remove *internal* inconsistencies of an otherwise satisfactory calculus. From the standpoint of the user, this accomplishes nothing unless he can also show that otherwise arbitrary postulates have any verifiable foundation in *external* experience. The engaging humour of his final remarks, which I shall now quote, suggest that Kendall himself is not wholly satisfied with the outcome of his pacific negotiations:

A friend of mine once remarked to me that if some people asserted that the earth rotated from east to west and others that it rotated from west to east, there would always be a few well-meaning citizens to suggest that perhaps there was something to be said for both sides, and that maybe it did a little of one and a little of the other; or that the truth probably lay between the extremes and perhaps it did not rotate at all.

It would be less necessary to insist that the issues at stake are of the utmost importance to the user, especially to the vast class of users who take the techniques on trust, if theoretical statisticians were content to arbitrate on the value of conclusions advanced by research workers on the basis of enquiries designed *ad hoc* and with due regard to background knowledge of the enquiry. Of late, more especially during the last fifteen years, they have in fact advanced claims with much wider terms of reference, as illustrated by the following citation from R. A. Fisher (*Design of Experiments*, 5th edn., pp. 7-9):

Inductive inference is the only process known to us by which essentially new knowledge comes into the world. To make clear the authentic conditions of its validity is the kind of contribution to the intellectual development of mankind which we should expect experimental science would ultimately supply. . . .

It is as well to remember in this connection that the principles and method of even *deductive* reasoning were probably unknown for several thousand years after the establishment of prosperous and cultured civilisations.

. . . The liberation of the human intellect must, however, remain incomplete so long as it is free only to work out the consequences of a prescribed body of dogmatic data, and is denied the access to unsuspected truths, which only direct observation can give. . . .

. . . The chapters which follow are designed to illustrate the principles which are *common to all experimentation*, by means of examples chosen for the simplicity with which these principles are brought out. Next, to exhibit the principal designs which have been found successful in that field of experimentation, namely agriculture, in which questions of design have been most thoroughly studied, and to illustrate their applicability to other fields of work. (*Italics inserted.*)

This passage is instructive more because of what it implies than because of what it explicitly asserts. We get the impression that recourse to statistical methods is prerequisite to the design of experiments of any sort whatever. In that event, the whole creation of experimental scientists from Gilbert and Hooke to J. J. Thomson and Morgan has been groaning and travailing in fruitless pain together; and the biologist of today has nothing to learn from well-tried methods which have led to the spectacular advances of the several branches of experimental science during the last three centuries. Nor is this all. We learn that we shall find the pattern of new and more powerful methods in a procedure for carrying out agricultural field trials prescribed, though Fisher does not say so, by one school of statisticians, and one alone.

What we then naturally ask is whether consequent advances of our knowledge of the soil, if any, have been commensurate with such a claim. In the parallel domain of marine biology, our knowledge of how to reproduce all relevant conditions for the culture of sea creatures has made great strides by recourse to entirely traditional principles of experimentation, while our theoretical knowledge of the growth needs of plants has not conspicuously broadened as the outcome of experiments designed in conformity with the demands of Greco-Latin Squares or randomised blocks. In the latter domain the claim that the theoretical statistician knows better than the man on the job how to do it is one which derives its sanction from a particular theory of statistical inference; and Fisher himself

would be first to admit this. If the theory turns out to be false, the result of increasingly widespread use of methods prescribed to design experiments must result both in curbing the ingenuity of the investigator at stupendous cost of time and in deterioration of standards of good workmanship in the laboratory. I believe we can already detect signs of such deterioration in the growing volume of published papers—especially in the domain of animal behaviour—recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration. Be that as it may, the fact that such methods are now in general use signifies that it is not merely an academic exercise to clarify the credentials of current views on statistical inference. Least of all is it merely a matter of moment to the trained mathematician at a time when trained mathematicians cannot reach agreement about them among themselves.

Should our adjudication lead us to embrace, with all its as yet half-formulated implications, the viewpoint of the new American school, the consequences will be far more drastic than many of our island contemporaries as yet recognise. In the closing words of his essay *Of the Academical or Sceptical Philosophy* Hume asks: “when we run over libraries persuaded of these principles, what havoc must we make?” Such havoc I suggest that little if anything in the cookery books will remain. We may have to reinstate statistics as continental demographers use the term. Laboratory experiments will have to stand on their own feet without protection from a façade of irrelevant computations. Sociologists will have to use their brains. In my view, science will not suffer.



PART I

---

*The Founding  
Fathers*



## CHAPTER TWO

# THE FOUNDING FATHERS AND THE NATURAL HISTORY OF GAMBLING

THE CURRENT DISPOSITION to regard stochastic\* considerations as prerequisite to the formulation of scientific laws derives its plausibility from two circumstances: (a) contemporary statistical theory embraces diverse domains of discourse into which the calculus of probability has intruded; (b) the laboratory worker too lightly assumes that there is general agreement about the relation of the calculus of probability to external events. The relation of the algebraic theory of probability to the real world is indeed more than ever before a topic of keen controversy. It therefore calls for critical re-examination at the outset; and we shall handicap ourselves unduly if we undertake our task on the optimistic assumption that all men of science are logically consistent. At the level of applied theory, we find in one and the same camp representatives of widely divergent views about when and in what sense the theory of probability can help us to interpret experience of the real world in a useful way. The core of the dispute is neither mathematical nor empirical. It is primarily a semantic issue which involves a fundamental dichotomy of attitudes with respect to the nature of valid judgment. As such it concerns us all. Were it not for the fact that most statisticians of middle age had already invested moral and intellectual capital in the Yule-Fisher tradition before existing differences became too acute to overlook, it would be hard to understand how British leaders of statistics, as cited in the last chapter, adroitly maintain an aspect of benevolent and seemingly nonchalant, though one may suspect uneasy, neutrality to the contestants.

Bacon somewhere speaks of man's inveterate habit of dwelling on abstractions. With equal propriety we may deplore a pernicious predilection of many highly intelligent people for *double-talk*. Rapid advances in the sciences of biology and chemistry after the mid-eighteenth century were in no small

\* See Chapter 5, p. 118.

measure made possible by a deliberate discipline to curb it. With experience of substantial progress during the preceding century behind them, men of science then had an object lesson before their eyes. The substitution of the word *gas* for *spirits* had exorcised a host of unclean superstitions at about the time when mathematics also annexed from common speech a word as redolent as spirits with emotive misconceptions. In seeking a relation between the calculus of *probability* and human action we thus encounter a difficulty which will dog our steps at every ensuing stage of our examination of the history of the topic. The word probability, like *bias*, *random*, *population*, *significance*, *likelihood*, *confidence*, and so many other terms in the vocabulary of statistics, carries over from common speech a miasma of irrelevant associations. While there is general agreement about the algebraic rules of the game, there is therefore still much controversy about the relevance of the rules to what we care to call *probability* in so far as the concept is referable to inference or to decision.

"Fundamentally the term probable," says Miss David (*Probability for Statistical Methods*, p. 1), "can only apply to the state of mind of the person who uses the word"; but she goes on to say that "the mathematical theory of probability is concerned, however, with building a bridge . . . between the sharply defined but artificial country of mathematical logic and the nebulous shadowy country of what is often termed the real world." In spite of her Platonic feelings about the real world, the ensuing discussion is essentially behaviouristic and one concludes that *fundamentally* in this context signifies "in everyday life." To a considerable school, however, *fundamentally* would here signify an article of faith which is through and through Platonic. To such, probability is a measure of the *legitimate intensity* of our conviction that a body of evidence justifies certain conclusions. This is the stand which Jeffreys and Carnap take. Kendall follows them when he speaks (Vol. I, p. 164) of the "attitudes of mind to which we relate the concept of probability."

The idealistic doctrine has a symbolism of its own (*vide infra* p. 51). Those who adopt it speak of: (i) the certainty that hypothesis *H* on data *q* is true as  $P(H/q) = 1$ ; (ii) the cer-

tainty that it is false on the same basis as  $P(H/q) = 0$ ; (iii) any other "state of mind" as  $P(H/q) = p$  such that  $0 < p < 1$ . If nothing we may infer from  $q$  has any bearing on  $H$ , proponents of the idealistic doctrine proceed in accordance with the axiom  $P(H/q) = \frac{1}{2}$  to build up a cumulative measure of conviction by iterative invocation of ignorance. Such is the *principle of insufficient reason*. To decide how contestants with different states of mind can agree about what particular value of  $p$  in a particular situation makes it a *legitimate* measure of our confidence in the truth of the hypothesis, we must either—as do Jeffreys and Carnap—fall back on such supposedly self-evident axioms or reinterpret our definitions in terms of observable occurrences.

If we follow the latter course, we shall speak only of the frequency with which our assertions consistent with a rule of assessing such and such evidence in such and such a way correspond with reality in the long run. It is in such strictly behaviourist terms that the experimentalist will presumably prefer to participate in the discussion of the relevance of the algebraic theory of probability to the real world. A sufficient objection to the alternative is the existence of a very large number of intelligent people, including not a few professional mathematicians, to whom the postulates invoked by the axiomatic school are by no means self-evident. In the commonly accepted sense of both terms, they are indeed amenable to proof or to disproof only if we can translate them into a frequency idiom referable to observable occurrences in contradistinction to inaccessible individual states of mind.

Happily, we do sometimes have the opportunity to observe mental processes to which a formal calculus of probability conceived in terms of the principle of insufficient reason, or in any other terms divorced from experience, should have a peculiar relevance if we have also good reason for subscribing to the faith of the axiomatic school. Levy and Roth (*Elements of Probability*, p. 14) cite an instructive example of such situations:

If 100 persons each have to choose a number between 0 and 9 inclusive, how often will the numbers 0, 1, 2 . . . be chosen? The abstraction which a mathematician *might* make from this problem

would leave him with a purely mathematical question concerning arrangements, the answer to which is, that each of the numbers will "probably" be chosen ten times. But this is not the real question; what we want to know is how people *actually choose*, and here we are faced by considerations of a psychological and social nature. In point of fact it has been found by actual testing of a large number of individuals that 7 and 3 are much more frequently chosen than any other number; these numbers both, of course, have a long historical and religious tradition behind them. As we see from such an example, the question whether the abstraction may be validly applied in a given case is not to be begged. The mathematical problem deals with the number of arrangements that can be conceived as possible in the circumstances, the physical problem with the groups of these which actually come into play. We can develop a mathematical theory of arrangements but a separate justification has to be found for it if it is to have practical applications. Thus, the mathematician may postulate that "an event can happen in two different ways"; whereas the physicist knows that it does happen in one way only.

The last sentence puts the spotlight on the main theme of this chapter. I here propose to deal with three questions in the following order:

(a) in what sense did the architects of the algebraic theory of probability themselves conceive it to be relevant to the real world?

(b) in what terms did they formulate the rules of the calculus and what latent assumptions about the relevance of the rules to reality do their axioms endorse?

(c) to what extent does experience confirm the factual relevance of the rules in the native domain of their application?

Our first question then is: *in what sense did the architects of the algebraic theory of probability themselves conceive it to be relevant to the real world?* This at least is answerable in non-controversial terms. An algebraic calculus of probability takes its origin from a correspondence between Pascal and Fermat over the fortunes and misfortunes of the Chevalier de Méré, a great gambler and by that token *très bon esprit*, but alas (wrote Pascal) *il n'est pas géomètre*. Alas indeed. The Chevalier had made his pile by always betting small favourable odds on getting at

least one six in 4 tosses of a die, and had then lost it by always betting small odds on getting at least one double six in 24 double tosses. Thus the problem out of which the calculus took shape was eminently practical, viz. *how to adjust the stakes in a game of chance in accordance with a rule which ensures success if applied consistently regardless of the fortunes of the session*. This is the theme song of the later treatise by James Bernoulli, and the major pre-occupation of all the writers of the classical period, including de Moivre, D. Bernoulli, d'Alembert and Euler.

Thus there is no ambiguity about what the Founding Fathers conceived to be the type specimen of real situations in which a calculus of probability is usefully applicable. Their formal statement of the fundamental operations of such a calculus is essentially identical with what is now current, though different schools of doctrine verbalise the initial assumptions in different terms. The end they had in view is beyond dispute; but the relevance of the assumptions implicit in the solution they offered is still highly debatable. Despite such obscurities not yet resolved by common consent, we may conceivably be satisfied that theory correctly prescribes a regimen of practice consistent with the intentions of writers in the classical period. If so, we may decently refrain from participation in the controversy without relinquishing the right to apply the rules of the calculus in strictly comparable situations. Our concern will then be to assess the claim that any one of the multitudinous applications of the calculus endorsed by the current theory of statistics is truly identifiable as a situation on all fours with the prescription of a reliable rule for division of stakes in a game of chance.

The ideological climate of our time is not propitious to success if we undertake the task with a light heart. Many of us, including the writer, have a puritan distaste for gambling and at best a superficial familiarity with the type of problems Pascal and his immediate successors tackled. At the outset, we should therefore be very clear about what we here mean by a betting rule. Let us accordingly scrutinise a situation analogous to the dilemma of the Chevalier through the eyes of his own generation. Our task will be to state a rule for division of stakes to ensure success to a gambler who bets on the outcome of

taking 5 cards *simultaneously* from a well-shuffled pack. We shall assume his bet to be that 3 of the cards will be pictures and that 2 will be aces. Without comment on the rationale of the rule, we shall first illustrate what the operations of the calculus are, and then how we prescribe the rule deemed to be consistent with the outcome. If we approach our problem in the mood of the Founding Fathers of the algebraic theory, we shall reason intuitively as follows.

In a full deck of 52 there are  $52^{(5)}$  ways\* in which the disposition of the cards *may* occur, this being the number of linear permutations of 52 objects taken 5 at a time without replacement. Out of these  $52^{(5)} = 311,853,201$  ways in which the cards might occur, the number of ways in which the disposition is consistent with the bet is  $5_{(2)} \cdot 4^{(2)} \cdot 12^{(3)} = 158,400$ , this being (*vide infra*, p. 41) the number of recognisably different linear permutations of 5 things taken from 52 when: (a) we regard 4 of the 52 as identical members of class Q and 12 as identical members of class P; (b) the sample consists of 2 members of Q and 3 members of P. Without here pausing to examine the relation between such usage and what probability signifies in everyday speech, we shall now arbitrarily define the *algebraic* probability of a *success*, i.e. of the specified 5-fold choice, and hence of the truth of the assertion the gambler proposes to make, in the classical manner, that is to say, the ratio  $5_{(2)} \cdot 4^{(2)} \cdot 12^{(3)} \div 52^{(5)} = 1 : 1,969$ .

To get the terms of reference of the operations of the calculus vividly into focus in its initial domain of practice before making them more explicit, we may then postulate fictitiously that *exactly* one in every 1,969 games justifies the gambler's assertion. In this fictitious set-up, we shall suppose that:

- (a) the gambler A bets that the result will be a success and agrees to pay on each occasion if wrong a penalty of £ $x$  to his opponent B;
- (b) his opponent likewise bets that the result will be a failure and agrees to pay to A a penalty of £ $y$  if wrong;
- (c) each gambler adheres to his system throughout a sequence of 1,969 withdrawals.

\* Here, as elsewhere, I adopt Aitken's economical symbolism:  $r_{(x)} = r! \div x! (r-x)!$  The symbol  $r^{(x)}$  has the usual meaning, i.e.  $r^{(x)} = r! \div (r-x)!$



If  $x = 1$  and  $y = 1,968$ , A will part with £1 on 1,968 occasions or £1,968 in all and B will part with £1,968 on only one occasion. Thus neither will lose or gain in any completed 1,969-fold sequence. If  $x = 1.05$ , so that A parts with a guinea whenever wrong and  $y = 1,968$  as before, A will part with £2,066 8s. od. in a complete 1,969-fold sequence and B will gain £98 8s. od. at his expense. If  $x = 1$  but  $y = 2,000$ , A will gain £32 in each such sequence and B will lose that amount. On the foregoing assumption, therefore, A will always gain if two conditions hold good:

(i) A consistently follows the rule of asserting that the result of a 5-fold withdrawal will be a success;

(ii) B agrees to pay a forfeit somewhat greater than £1,968 when A is right, on the understanding that A agrees to pay a forfeit of £1 if wrong.

Needless to say, the Founding Fathers did *not* postulate that the gambler A would score exactly one success in every 1,969-fold sequence. What they did claim is that the ratio of success to failure, i.e. of true to false assertion, would be 1 : 1,968, if he went on making the same bet consistently in a sufficiently extended succession of games. Such then are the odds in favour of success in the idiom of the game. In so far as the classical formulation has any bearing on a calculus of truth and falsehood, it thus refers to the long-run frequency of correct assertion in conformity with a rule *stated in advance*. We are then *looking forwards*. Nothing we observe on any single occasion entitles us to deviate from adherence to it. Nothing we claim for the usefulness of the calculus confers a numerical valuation on a *retrospective judgment* referable to information derived from observing the outcome of a particular game.

In stating the foregoing rule and its application, we have identified the long-run ratio of the number of successes which *will* occur to the number of failures which *will* also occur with the ratio of the number of different ways in which success *may* occur to the number of different ways in which failure *may* occur. The italicised auxiliaries suffice to show that this calls for justification; but the proponents of the classical theory seem to have been quite satisfied to embrace the identification

as a self-evident principle. Even so, we shall see that it does not suffice to prescribe the circumstances in which the calculus actually specifies long-term experience of situations to which they applied it, or even to prescribe any class of situations to which the rule might conceivably be relevant. This will emerge more clearly if we now formulate the definition of the concept and the rules it endorses more explicitly.

We thus come to the second question we have set ourselves to answer: *in what terms did the Founding Fathers formulate the rules of the calculus and what latent assumptions about the relevance of the rules to reality do the axioms condone?* To do justice to this, we shall first need to distinguish between two ways in which we may choose 5 cards: (a) *exhaustively*, i.e. simultaneously or successively without replacement of any one of the 5 cards chosen before picking its successor; (b) *repetitively*, i.e. picking the cards successively and *replacing* each card after recording its denomination before picking a successor, or cutting successively and recording each time the bottom card exposed before reassembling the pack.

If we now speak provisionally of our card pack in more general terms as the *universe of choice*, we may formally state as follows a comprehensive algebraic definition of probability consistent with the classical approach in the discrete domain of the game of chance:

If  $x_j$  or  $x_{(j)}$  denote the number of different linear permutations of  $r$  items from an  $n$ -fold universe consistent with the specification of the way of choosing a sample of class  $J$  and  $x_{r,n}$  or  $x_{(r,n)}$  is the corresponding number of all different linear  $r$ -fold permutations from the same  $n$ -fold universe in conformity with the *same rule of exclusion*, the probability of observing the specified sample class is the ratio of  $x_j$  :  $x_{r,n}$  or  $x_{(j)}$  :  $x_{(r,n)}$ .

The insertion of the words italicised leaves open the alternative prescription of sampling exhaustively (i.e. *without* replacement) or repetitively (i.e. *with* replacement as defined); and the use of the brackets in  $x_{(j)}$  and  $x_{(r,n)}$  signifies that sampling occurs without replacement. Such a definition subsumes all the elementary rules of the calculus of probability in what Macmahon calls the *Master Theorem* of permutations. This is

as follows. If an  $n$ -fold universe is exhaustively and exclusively classifiable as  $a$  items of class A,  $b$  of class B,  $c$  of class C, etc., the number of  $r$ -fold linear permutations containing  $u$  items of class A,  $v$  of class B,  $w$  of class C, etc., is:

*with replacement:*

$$x_j = \frac{r!}{u! v! w! \dots} a^u \cdot b^v \cdot c^w \dots \quad (i)$$

*without replacement:*

$$x_{(j)} = \frac{r!}{u! v! w! \dots} a^{(u)} \cdot b^{(v)} \cdot c^{(w)} \dots \quad (ii)$$

The alternative assumptions w.r.t. the rule of exclusion respectively prescribe  $x_{r,n} = n^r$  and  $x_{(r,n)} = n^{(r)}$ , whence the appropriate probabilities referable to the  $r$ -fold sample of type J defined as such by  $a, b, c, \dots u, v, w$ , etc., are:

*with replacement:*

$$P_j = \frac{r!}{u! v! w! \dots} \left(\frac{a}{n}\right)^u \left(\frac{b}{n}\right)^v \left(\frac{c}{n}\right)^w \dots \quad (iii)$$

*without replacement:*

$$P_{(j)} = \frac{r!}{u! v! w! \dots} \frac{a^{(u)} b^{(v)} c^{(w)} \dots}{n^{(r)}} \quad (iv)$$

These expressions respectively correspond to the general term of the multinomial theorem in its customary form and in the form of which Vandermonde's theorem is a special case. Each defines the distribution law appropriate to the type of sampling specified, and each subsumes the two fundamental laws of the calculus. It will suffice to state the latter in the simplest form as below, i.e. for two score values, whence the more general statement follows by iteration:

*The Multiplicative Property.* If  $p_b$  is the unconditional probability of first recording a score  $b$  and  $p_{c.b}$  is the conditional probability of then recording a score  $c$ , the joint probability of recording both  $b$  and  $c$  in that order is  $p_{bc} = p_b \cdot p_{c.b}$ .

*The Additive Property.* If  $b$  and  $c$  are among possible score values, only one of which one may record on any one occasion, the

probability that the score will be either  $b$  or  $c$  on such an occasion is  $P(b \text{ or } c) = p_b + p_c$ .

A single illustration will suffice to show that (iii) and (iv) do indeed subsume both rules. The probability of the compound choice  $J$  specified as the extraction of 2 spades and 1 red card in a 3-fold exhaustive (with removal) trial is  ${}_{3(2)} \cdot {}_{13}^{(2)} \cdot {}_{26}^{(1)} \div {}_{52}^{(3)}$ . This is also the probability of getting any one of the sequences  $SSR$ ,  $SRS$  or  $RSS$ , and we may write in the foregoing notation:

$$\begin{aligned} \frac{{}_{3(2)} \cdot {}_{13}^{(2)} \cdot {}_{26}^{(1)}}{{}_{52}^{(3)}} &= \frac{{}_{13} \cdot {}_{12} \cdot {}_{26}}{52 \cdot 51 \cdot 50} + \frac{{}_{13} \cdot {}_{26} \cdot {}_{12}}{52 \cdot 51 \cdot 50} + \frac{{}_{26} \cdot {}_{13} \cdot {}_{12}}{52 \cdot 51 \cdot 50} \\ &= p_s \cdot p_{s,s} \cdot p_{r,ss} + p_s \cdot p_{r,s} \cdot p_{s,sr} + p_r \cdot p_{s,r} \cdot p_{s,rs} \\ &= p_{ss} \cdot p_{r,ss} + p_{sr} \cdot p_{s,sr} + p_{rs} \cdot p_{s,rs} \end{aligned}$$

When sampling is repetitive (with replacement or without removal),  $p_{s,s} = p_s = p_{s,ss}$ , etc., and the foregoing reduces to

$${}_{3(2)} \cdot \left(\frac{{}_{13}}{52}\right)^2 \left(\frac{{}_{26}}{52}\right) = p_s^2 \cdot p_r + p_s \cdot p_r \cdot p_s + p_r \cdot p_s^2 = 3p_s^2 \cdot p_r$$

To exhibit these expressions as terms of the appropriate multinomial, we recall that spades and reds together with clubs constitute an exclusive system, as follows:

<i>Spades</i>	<i>Red</i>	<i>Clubs</i>	<i>Total</i>
${}_{13}$	${}_{26}$	${}_{13}$	$n = 52$
$2$	$1$	$0$	$r = 3$

In full, since  ${}_{13}^{(0)} = 1 = {}_{13}^{(0)}$  and  $0! = 1$ :

$$\begin{aligned} {}_{3(2)} \cdot \frac{{}_{13}^{(2)} \cdot {}_{26}^{(1)}}{{}_{52}^{(3)}} &= \frac{3!}{2! 1! 0!} \frac{{}_{13}^{(2)} \cdot {}_{26}^{(1)} \cdot {}_{13}^{(0)}}{{}_{52}^{(3)}} \\ {}_{3(2)} \cdot \left(\frac{{}_{13}}{52}\right)^2 \cdot \frac{{}_{26}}{52} &= \frac{3!}{2! 1! 0!} \left(\frac{{}_{13}}{52}\right)^2 \left(\frac{{}_{26}}{52}\right)^1 \left(\frac{{}_{13}}{52}\right)^0 \end{aligned}$$

These are respectively the appropriate terms of:

$$({}_{13} + {}_{26} + {}_{13})^{(3)} \div {}_{52}^{(3)}$$

and

$$\left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4}\right)^3 = ({}_{13} + {}_{26} + {}_{13})^3 \div 52^3$$

Actually, (iii) and (iv) subsume the rules of the calculus in a more comprehensive way than any formulation of the classical period, here defined as that which antedates the writings of Laplace; but the idiom of choice in the foregoing definition is wholly consistent with the usage of the Founding Fathers. The terms used call for careful scrutiny, if we wish to be clear about their relevance to the frequency of success in games of chance. At the outset, we have identified *all possible ways* of choosing 5 cards specified in a particular way with all *linear permutations* consistent with the specification, i.e. referable to a particular combination. This identification of our unit of choice with a single linear arrangement will prove to be highly suggestive, when we later explore circumstances in which our formal definition is itself identifiable as a frequency ratio; but a blunder of D'Alembert in his article on probability in the first edition of the *Encyclopédie* reminds us how liberally writers of the classical period relied on their own intuitions to assign to the combination and to its constituent permutations their agreed role in the theory.

\*                      \*                      \*                      \*

In the classical context of the gaming table, the factual application of the rules of the calculus presupposes the existence of: (a) an *agent*, the player and/or dealer; (b) an *apparatus*, card pack, die or lottery; (c) a *programme of instructions*; (d) a *wager*, i.e. an unchangeable *assertion* of the outcome of each trial. If the apparatus is a card pack or lottery urn, the programme of instruction will include a specification of the act of choice in the most literal sense of the term. One difficulty which confronts a confident evaluation of the viewpoint of the Founding Fathers arises from the circumstance that verbal specification of different methods of choice in textbook expositions of combinations and permutations leaves much to the imagination of the reader. Before trying to get to grips with a theory of chance which relies on a calculus of choice, it may therefore be profitable to sharpen the outline of situations in which expressions for the enumeration of all possible acts of choice consistent with a particular prescription are also in fact consistent with the range of realisable possibilities. The

reader who regards the undertaking as trivial may skip what precedes the next row of asterisks.

At this stage, we shall refrain from asking whether such considerations have any bearing on how *often* the gambler actually chooses a sample of one or other class specified as a particular combination in this context. The word *actually* as used by Levy and Roth in the passage cited above has indeed other implications worthy of notice as a qualifier of the verb which follows it. For specification of the particular *unit of choice* in our foregoing definition imposes unstated factual restrictions on any conceivable relevance of probability so defined to what we may literally or metaphorically mean by choice in successive games of chance. We may usefully examine such restrictions in this context because it is easy to overlook them when theory invokes the sanction of the classical doctrine to identify the hypothetical infinite population of Laplace with what we here refer to provisionally, and somewhat ambiguously, as the universe of choice. The meaning we can rightly assign to our universe of choice, and the sense in which we can usefully conceive it as finite or otherwise, depends on the system of choice we prescribe.

In the formal calculus of choice we mean by all linear permutations of 5 cards of a particular specification all the different ways in which we can lay them out face upwards in a line, if free to do anything we like with a single pack or with a sufficient number of identical packs but with due regard to the explicit alternative conditions customarily distinguished as *with* and *without* replacement. Choice *without* replacement then merely means that no card of a particular denomination may be present more than once in one and the same sequence. Choice with replacement signifies that cards of one and the same denomination may occur any number of times from 0 to  $r$  inclusive in one and the same  $r$ -fold sequence. Though this broad distinction between sampling without and sampling with replacement suffices to specify appropriately all possible ways of choosing our cards when we are free to handle the packs face upwards, it emphatically does not suffice to specify appropriately all possible ways of doing so when we are choosing cards face downwards from one or even from an

indefinitely large number of identical packs. We then have to prescribe factual safeguards which the foregoing definition of *algebraic* probability does not make explicit.

Whether or no our definition has any justifiable relevance to the frequency of success in games of chance, it certainly cannot claim to have any relevance to a succession of games in which the possibilities it specifies are *not* realisable. Clearly, therefore, some safeguards must be explicit if our calculus is to describe what the gambler *actually* chooses. If we specify all different linear permutations of cards in a pack or sample sequences consistent with the criterion of success as information relevant to long-term experience of sampling in conformity with the prescription of such a game, we imply the actuality of making their acquaintance if we go on playing the same game long enough. That this need not be so, unless we prescribe the rules appropriately, will be sufficiently evident if we distinguish as follows between two different ways of actually choosing blindfold *without* replacement:

- (i) we take each card of the  $r$ -fold pack without restriction from any part of the pack;
- (ii) we extract them simultaneously, i.e. we pick out *with one hand* an  $r$ -fold sequence of cards lying consecutively in the deck *held by the other*.

If we adopt the latter procedure, there will be only  $(n-r+1)$  ways of selecting an  $r$ -fold sample from one and the same finite  $n$ -fold universe identified as such with any particular pack. Thus there will be 48 different ways of choosing 5 cards, and none of these will include 2 aces and 3 pictures if we pick the 5 from the middle of a pack with all aces at the bottom of the deck or all the pictures at the top. If we always use a new pack stacked in the same way, we shall therefore never meet more than 48 different 5-fold sequences. We can hope eventually to meet all the possible ways implicit in our initial algebraic definition of probability only if we *shuffle* the same pack before each game or use different packs stacked in every possible order. Either way, what we have here provisionally called *the* universe of choice changes from sample to sample, and the definite article is justifiable only in so far as the com-

position of any full pack specifies all the *numerical* information relevant to our definition.

If we adopt the alternative stated above, we are free to choose each card of the  $r$ -fold sample from any part of the pack. On the assumption that we fully exploit this freedom to choose each card from *any* part of the pack, we may eventually hope to meet all possible  $n^{(r)}$  different  $r$ -fold samples of exhaustive choice and all  $x_{(j)}$  different  $r$ -fold samples of the particular class J in successive  $r$ -fold withdrawals from  $n$ -fold card packs stacked in exactly the same way and hence identical in all respects. None the less, we cannot speak with propriety of sampling without replacement from a finite universe with the implication that we are talking about one and the same card pack throughout. Neither sampling successively without replacement nor sampling simultaneously is consistent with the identification of the universe of choice with a single particular card pack.

When we speak of choosing cards from a pack, we do not necessarily imply that we remove them. If the cards lie face downwards, choosing *with replacement* subsumes any method of choice which leaves us free to record any card of a particular denomination from 0 to  $r$  times. For instance, we may cut the pack and turn the top set face upwards. The presumption then is that we might equally well entrust the task to a dealer. Thus we need not interpret the term choice in the most literal way. To say that we may choose  $52^5$  different sequences of 5 cards if allowed to record the result of cutting 5 times anywhere or of picking out 5 cards in succession, replacing each one except the last before taking the next, merely states that there are  $52^5$  different ways in which *anyone* may arrange 5 cards, face upwards in a row, if free to take one card from *any* part of 5 identical full packs, and hence  $52^5$  distinguishable hands which we might receive from the dealer who picked out a single card from *any* position in each deck of cards. On the other hand, there would be only one possible way of dealing the 5 cards singly from the 5 packs, if the dealer took each card from the top or from any preassigned level in the deck in an assigned order. If we therefore assert that all possible ways of *choosing* with replacement 5 cards from a full pack correspond to all



possible ways of *dealing* one card from each of 5 packs, we implicitly impose some plan of action on the dealer.

More than one such plan is consistent with the same formal statement of the choice. This will be clear if we think of 6 packs each consisting of the ace, 2, 3, 4, 5 and 6 of spades. If we score a sequence by adding up the number of pips on the constituent cards and distinguish a class of 3 selections by the fact that the score is 5, the class itself subsumes 6 different sequences:

$$113, 131, 311 ; 122, 212, 221$$

In the same set-up all possible 3-fold sequences that we may distinguish are  $6^3 = 216$ . If we speak accordingly of  $6 \div 6^3 = \frac{1}{36}$  as the *probability* of scoring 5 in sampling with replacement from the 6-fold pack, the classical prescription embraces several factually different situations:

(i) the player or the dealer removes 3 cards singly in succession from *anywhere* in the same pack and replaces each card *anywhere* after recording the scores before drawing another;

(ii) the player or the dealer cuts 3 times the same pack *anywhere* replacing the cards in their original position after recording the exposed card;

(iii) the player or dealer cuts *anywhere* each of 3 identical packs recording each result and replacing the cards in their original order;

(iv) the player or dealer removes a single card from *anywhere* in each of 3 packs, recording the result and replacing each card *anywhere* in its own pack.

Inasmuch as we have defined probability with reference to all possible sequences we may encounter, the word *anywhere* in this context has therefore a special meaning. It endorses the possibility that we may eventually meet all possible sequences, the implication being that we follow no set plan w.r.t. the level from which we take a card, the level at which we replace it, or the level at which we cut the pack, in a sufficiently protracted succession of games. It is irrelevant to make this explicit if we alternatively prescribe:

(a) a *reshuffle* of the pack of (i) or (ii) between taking from it an individual card and replacing it;

(b) a reshuffle of each of the 3 packs in (iii) and (iv) between successive games.

What we thus mean by choice with replacement in the face-upward setting of the algebra of permutations and combinations embraces a diversity of factual situations in which all possible ways of actually choosing a particular hand in a game of chance will not necessarily correspond to all possible ways prescribed by a formal calculus of choice, unless our programme of instructions prescribes how we actually choose them. How *often* we actually choose them is another matter which will be the theme of Chapter 3. Nothing we have said so far suffices to justify the conviction that the ratio of all different ways of scoring a factually realisable success to all different methods of scoring either a success or a failure on the same assumption will be the same as the ratio of actual successes to successes and failures in a long enough sequence of games.

On this understanding, we are free to extend our definition to games of chance other than card games. By putting the dealer into the picture, we have absolved ourselves from the need to interpret the term choice in its most usual sense. The score the player records when the dealer, as the *agent*, cuts the pack is in no sense more literally an act of choice than the score the player records when the cubical die comes to rest on the floor. The specification of all different ways in which a cubical die may lie uppermost in a 3-fold toss and all different ways in which we may score 5 in a 3-fold toss is thus formally identical with the foregoing specification for the deck of six spades, if we also postulate that each face will at some time lie uppermost in a sufficiently prolonged succession of tosses; but we have not absolved ourselves from the need to make explicit the circumstances in which we can confidently assert that it will do so, except in so far as we assume that the act of tossing is comparable to cutting *anywhere*. We can then speak of the die as an unchanging finite universe of choice in the sense that the pack from which we cut *anywhere* is an unchanging (finite) universe of choice if the dealer restores the deck to its original arrangement after recording the cut; but we have still

to satisfy ourselves that the dealer's utmost efforts will suffice to justify the rule for the division of the stakes. We shall later see that the manufacturer of the die must share some of the responsibility with the dealer.

\*                      \*                      \*                      \*

The foregoing digression to clarify the factual content of the term *choice* in the context of the game of chance will not have been fruitless, if it forces us to recognise that an alternative formulation of the basic concept, now more fashionable than the foregoing, itself discloses no explicit indications of circumstances which endorse the factual relevance of the rules of the calculus to the prescription of a reliable betting rule, still less to the wider terrain of situations annexed by contemporary statistical theory. The restatement of the classical doctrine in the symbolism of the modern theory of sets is indeed merely a refinement of the classical notation. It owes its appeal to the mathematician because the theory of sets can accommodate points as well as discrete objects and hence annexes the continuum as its legitimate territory.

In the discrete domain of the game of chance, we think of a set  $S$  as a collection of  $n(S)$  items with an attribute  $A_s$  which each of its members possesses, and say that an object  $O$  is in  $S$  if it has the attribute  $A_s$ . An object may be in more than one set, and we say that it is in the set  $PQ$  of  $n(PQ)$  items if it has  $A_p$  and  $A_q$ . We say that it is in the set  $(Q + P) = (P + Q)$  if it has  $A_q$  without  $A_p$ ,  $A_p$  without  $A_q$  or  $A_p$  with  $A_q$ , i.e. is in the set made up of the sets  $P$  and  $Q$  alone, whence  $n(P + Q) = n(P) + n(Q) - n(PQ)$ . If all members of  $P$  are members of  $S$ , we speak of  $P$  as a subset of  $S$ , and of  $S$  as the *universal set* if it includes every other set in the field of discourse as a subset. Thus the set  $H$  of all hearts is a subset of the universal set  $F$  of the full pack, as is also the subset  $E$  of all cards with an even number of pips. The set  $HE = EH$  of all hearts with an even number of pips is a subset of the set  $(E + H)$  which includes all hearts, all cards (including hearts) with an even number of pips and no others, and the set  $(E + H)$  is itself a subset of  $F$ . Thus  $n(F) = 52$ ,  $n(E) = 20$ ,  $n(H) = 13$ ,  $n(EH) = 5$ ,  $n(E + H) = 13 + 15 = 28 = n(E) + n(H) - n(EH)$ . Those

who use the set theory symbolism define the probability that O in the set M has  $A_q$  as

$$P(Q/M) = n(MQ) \div n(M).$$

Whence we get for O in F interpreted as above

$$P(\overline{E + H}/F) = P(E/F) + P(H/F) - P(HE/F) \quad (v)$$

$$P(HE/E) \cdot P(E/F) = P(HE/F) = P(HE/H) \cdot P(H/F) \quad (vi)$$

The above are the two fundamental theorems of addition and multiplication each in its most general form for 2 criteria of classification of the event. They are easily adaptable to accommodate more than 2 criteria by simple iteration. Thus the reader should be able to derive:

$$\begin{aligned} P(\overline{A + B + C}/S) &= P(A/S) + P(B/S) + P(C/S) \\ &\quad - P(AB/S) - P(AC/S) - P(BC/S) + P(ABC/S) \end{aligned}$$

In this idiom, we say that two sets H and K are *exclusive* if no object is in both, so that  $n(HK) = 0$  and  $P(\overline{H + K}/F) = P(H/F) + P(K/F)$ . Thus if K is the set of black aces,  $n(K) = 2$ ,  $n(H + K) = 15$ ,  $n(HK) = 0$  and  $P(H + K/F) = 15/52$ . We say that two subsets H and K of F are *independent* if  $n(HK) \div n(K) = n(H) \div n(F)$  and  $n(HK) \div n(H) = n(K) \div n(F)$ . Thus if  $K = E$ , so that  $n(KH) = 5$ ,  $n(H) = 13$ ,  $n(E) = 20$  and  $n(F) = 52$ , we have  $P(HE/F) = P(H/F) \cdot P(E/F)$ . In this more restricted form, the rule of addition is implicit in (iii) and (iv) above.

In the foregoing example, we have defined an object illustratively as a single card from the full pack which is the universal set, called also the *fundamental probability set* (F.P.S.). Thus O is a unit sample and F is the universe of choice. The fundamental rules of the calculus embodied in (v) and (vi) still hold good if we define an object as an  $r$ -fold sample for  $r > 1$ , and the F.P.S. as the set of all different  $r$ -fold samples distinguishable in terms of the relevant criteria of classification applicable to each constituent event and the sequence of events so distinguished. We thus get back to the classical definition in terms of linear permutations. All we have achieved is that we have made the addition rule in its most general form explicit

at the outset with a formal endorsement to extend the terms of reference of the calculus into the continuum.

The explicitly idealistic symbolism mentioned above (p. 34) is an adaptation of the foregoing. Those who use it, define the assertion *O in Q is also in M*, i.e. *O* with the attribute  $A_q$  has also the attribute  $A_m$ , as the hypothesis  $H_m$  on data  $q$ . They then write  $P(M/Q) = P(H_m/q)$ . In either domain of symbolism we assign the probability  $\frac{1}{2}$  to heads in single tosses of a penny because  $n(M) = \frac{1}{2}n(Q)$  if  $A_q$  characterises the set of all properly minted pennies after a toss and  $A_m$  the set of all pennies which *supposedly* lie heads uppermost. When they assign this value some proponents of the set-theory definition, e.g. Neyman, seemingly do so (*vide infra* p. 55) on the understanding that the identity stated suffices to endorse a *prospective* rule about betting on the outcome of a long sequence of tosses; but those who use the  $H./q$  symbolism make no such prospective claim.

As I understand it, their convention extends to the following situation. A normal penny lies under the mat. Before lifting the mat I know nothing about the penny other than that it has 2 faces, i.e.  $n(F) = 2$ , one only classified as a tail, i.e.  $n(TF) = 1$ . Since  $n(TF) = \frac{1}{2}n(F)$ , on these data " $q$ " we assign to the hypothesis  $H$ , viz. to the value of the *retrospective* verdict that the unseen penny lies tail uppermost, the probability  $P(H_i/q) = \frac{1}{2}$ ; but what we mean by probability in this setting is equally consistent with the assertion that the penny fell on its face after a spin or that someone placed it tails up before placing the mat on top of it. If we therefore try to translate our symbols into the public idiom of frequency, we merely get back to admission of ignorance. It may well be that I here misinterpret what Jeffreys or Carnap conceive as the proper interpretation of this  $H.q$  notation. If so, I must plead that the numerical assessment of states of ignorance about numismatic data at the level of individual judgment conveys nothing intelligible to me.

It will now be clear that the set-theory restatement of the calculus embraces exactly the same set of operations as the classical statement, as does also the explicitly idealistic interpretation of the initial concepts in the restatement of Jeffreys or Carnap. Different conventions of symbolism are current and

different interpretations of the symbols at the verbal level endorse different types of statement about their conceivable relevance to practical affairs. What we mean by the *operations* of a calculus of probability thus admits no difference of opinion; but we are as yet no nearer to understanding the factual relevance of the rules to the class of situations in which the Founding Fathers invoked them. No algebraic definition of probability so far stated in terms of what *may* occur or of what *has* occurred endorses an indisputable warrant for a rule of conduct conceived in terms of how *often* it *will* occur.

The reader may therefore forgivably feel that an answer to our third question is overdue: *to what extent does experience endorse the factual relevance of the rules in the original domain of their application?* Let us admit that few of us seriously doubt the reality of some factual basis for the faith of the Founding Fathers, and the relevance of their faith to the fate and fortunes of the Chevalier. A few investigations recorded in more recent times will indeed encourage us still to explore hopefully the credentials of a doctrine with so little ostensible relevance to its applications at the most elementary level. All we shall ask of the theory at this stage is that it *works*, if the gambler goes on *long* enough. We here interpret this to mean that:

(a) over 1,000 is a *large enough* number of games in the context;

(b) as a criterion of what *works*, we shall be satisfied with a correspondence between theory and observation if it conforms to the familiar standard set by the tabular content of any current textbook of physical chemistry.

Karl Pearson (*Biometrika* 16) cites the following counts for 3,400 hands (of 13 cards) at whist with corresponding probabilities calculated in accordance with the distribution defined by the terms of  $(13 + 39)^{(13)} \div 52^{(13)}$ :

<i>No. of Trumps per hand</i>	<i>No. of hands observed</i>	<i>No. of hands expected</i>
<i>under 3</i>	1,021	1,016
<i>3-4</i>	1,788	1,785
<i>over 4</i>	591	599

Uspensky (*Introduction to Mathematical Probability*) records two

experiments of this type. The first records 7 experiments each based on 1,000 games in which the score is a success if a card of each suit occurs in a 4-fold simultaneous withdrawal from a pack without picture cards. The probability of success is thus:

$$\frac{4!}{1! 1! 1! 1!} \frac{10^{(1)} 10^{(1)} 10^{(1)} 10^{(1)}}{40^{(4)}} = 0.1094$$

The outcome for the seven successive 1,000-fold trials (I-VII) was as follows:

I	II	III	IV	V	VI	VII
0.113	0.113	0.103	0.105	0.105	0.118	0.108

A second experiment subsumes 1,000 games in which the score for the 5-fold withdrawal from a full pack is the number of different denominations cited below with the corresponding theoretical and observed proportions correct to 3 places:

	1+1+1+1+1	2+1+1+1	2+2+1
Observed	0.503	0.436	0.045
Expected	0.507	0.423	0.048
	3+1+1	3+2	4+1
Observed	0.014	0.002	0.000
Expected	0.021	0.001	0.000

Many experiments purporting to vindicate the calculus merely show that the mean proportion of successes in large trials lies close to what the classical writers would call the proportion of favourable cases, viz. 0.5, if we score each head as a success in a long sequence of coin tosses. Good agreement then merely shows that assumptions consistent with a naïve adherence to the set theory definition of the probability of success in a *single* trial are acceptable. For reasons discussed below (p. 62), they do not necessarily vindicate the claim of the calculus to prescribe the long run frequency of the *r*-fold trial for values of *r* other than unity.

In the domain of die models, results obtained from experiments on the *needle problem* of the eighteenth-century French naturalist Buffon are both relevant and arresting. The gamester drops a needle of length *l* on a flat surface ruled with parallel

lines at a distance ( $h$ ) apart, scoring a success if it falls across and a failure if it falls between them. Theory prescribes that the ratio of success to failure involves  $\pi$ , the probability of success being  $2l/h\pi$ . If we equate this to the proportionate frequency of success, we may thus be able to give a more or less satisfactory evaluation for  $\pi$ , and its correspondence with the known value will then be a criterion of the adequacy of the calculus. Uspensky (*loc. cit.*) cites two such:

<i>Investigator</i>	<i>No. of throws</i>	<i>Estimate of <math>\pi</math></i>	<i>Error</i>
Wolf 1849-53	5,000	3·1596	<0·019
Smith 1855	3,204	3·1412-3·155	<0·015

The alternative figures in the second line of the table take doubtful intersections into account. In 1901, an Italian mathematician (cited by Kasner and Newman, *Mathematics and the Imagination*) carried out an experiment involving 3,408 tosses of the needle. Lazzerini obtained the value  $\pi \simeq 3·1415929$ , an error of only 0·0000003.

The theory of this experiment need not concern us here; but the nature of the game calls for comment because the model set-up is seemingly more comparable to that of the Chevalier's die than are some situations cited above. In the context elsewhere cited Uspensky refers to records of Bancroft Brown on experience of American dice in the game of *craps*. The caster wins if he scores a *natural* (7 or 11) at the initial double toss, loses if he scores *craps* (2, 3 or 12) and otherwise has the right to toss his two dice till the score is the same on 2 successive occasions, in which event he wins, or till the score for the double toss is 7, in which event he loses. Correct to 3 places, the probability of winning is 0·493 and of scoring craps is 0·111. In 9,900 games recorded by Brown the corresponding frequencies were 0·492 and 0·106.

This is admittedly impressive. None the less, much more prolonged trials summarised by Keynes (*vide infra*) have dispelled the belief that a calculus conceived in the foregoing terms is consistent with any reasonable expectation of an equally spectacular correspondence between theory and the behaviour of European dice in commercial production at an earlier date. Admittedly, the discrepancies are not formidable



in terms of a division of stakes congenial to the gaming companions of *de Meré* in their cups; but we may dismiss the objection that they are attributable to lack of perseverance. Unless we equate the word *ordinary* to twentieth-century American, Neyman has therefore chosen a singularly unsatisfactory exhibit to vindicate the adequacy of the set theory formulation, when he declares in his recent *Introduction to Probability and Statistics* (p. 16):

An ordinary die has six sides. Hence the F.P.S. is composed of  $n(A) = 6$  elements. Only one of the sides has six dots on it. Hence  $n(AB) = 1$ . Hence  $P_1 \dots = \frac{1}{6}$ .

A customary rejoinder to objections of this sort is that the definition refers to an *unbiased* die; but this is merely abuse of the plaintiff's counsel if it carries with it no specification of how to construct a die which has no bias, i.e. one which does in fact conform to the rules of the calculus. In Neyman's system the F.P.S. of the penny has 2 elements, and the probability of scoring  $r$  heads in an  $r$ -fold toss is  $\frac{1}{2}$ . If it happens that 10,000 tosses yield a head score of 4,800, we might plausibly guess that our penny will subsequently behave in close accord with the long-run outcome of cutting cards from a 100-fold pack of which 48 were red and 52 black; but we should need experience of an endless series of trials to specify precisely the F.P.S. of our second-order model in the appropriate way. We are thus no nearer to our goal, i.e. how to prescribe the factual conditions relevant to the validity of a rule we have already undertaken to state *in advance*.

In a comprehensive survey of factual investigations into what actually happens in games of chance, Keynes draws attention to the delusion that the outcome has a special relevance to J. Bernoulli's *deductive* theorem (p. 86) of large numbers. All they can in fact do is to reinforce confidence in the empirical principle of statistical equilibrium in large-scale trials and exhibit how far the classical statement of the theory suffices to state in advance a reliable rule for large-scale operations. "We can seldom be certain," writes Keynes, "that the conditions assumed in Bernoulli's Theorem are fulfilled . . . the theorem predicts not what will happen but only what is, on

certain evidence, likely to happen. Thus even where our results do not verify Bernoulli's Theorem, the theorem is not thereby discredited." Keynes thus describes the outcome of experiments on die models.

The earliest recorded experiment was carried out by Buffon, who, assisted by a child tossing a coin into the air, played 2,048 *partis* of the Petersburg game, in which a coin is thrown successively until the *parti* is brought to an end by the appearance of heads. The same experiment was repeated by a young pupil of De Morgan's "for his own satisfaction." In Buffon's trials there were 1,992 tails to 2,048 heads; in Mr. H.'s (De Morgan's pupil) 2,044 tails to 2,048 heads. . . . Following in this same tradition is the experiment of Jevons, who made 2,048 throws of ten coins at a time, recording the proportion of heads at each throw and the proportion of heads altogether. In the whole number of 20,480 single throws, he obtained heads 10,353 times. . . . All these experiments, however, are thrown completely into the shade by the enormously extensive investigations of the Swiss astronomer Wolf, the earliest of which were published in 1850 and the latest in 1893. In his first set of experiments Wolf completed 1,000 sets of tosses with two dice, each set continuing until every one of the 21 possible combinations had occurred at least once. This involved altogether 97,899 tosses, and he then completed a total of 100,000. These *data* enabled him to work out a great number of calculations, of which Czuber quotes the following, namely a proportion of .83533 of unlike pairs, as against the theoretical value .83333, i.e.  $\frac{8}{9}$ . In his second set of experiments Wolf used two dice, one white and one red (in the first set the dice were indistinguishable), and completed 20,000 tosses. . . . He studied particularly the number of sequences with each die, and the relative frequency of each of the 36 possible combinations of the two dice. The sequences were somewhat fewer than they ought to have been, and the relative frequency of the different combinations very different indeed from what theory would predict. The explanation of this is easily found; for the records of the relative frequency of each face show that the dice must have been very irregular, the six face of the white die, for example, falling 38 per cent more often than the four face of the same die. This, then, is the sole conclusion of these immensely laborious experiments—that Wolf's dice were very ill made. . . . But ten years later Wolf embarked upon one more series of experiments, using *four* distinguishable dice—white, yellow, red, and blue—and tossing this set of four 10,000 times.

Wolf recorded altogether, therefore, in the course of his life 280,000 results of tossing individual dice. It is not clear that Wolf had any well-defined object in view in making these records, which are published in curious conjunction with various astronomical results, and they afford a wonderful example of the pure love of experiment and observation.

Of lotteries Keynes records the following particulars:

Czuber has made calculations based on the lotteries of Prague (2,854 drawings) and Brunn (2,703 drawings) between the years 1754 and 1886, in which the actual results agree very well with theoretical predictions. Fechner employed the lists of the ten State lotteries of Saxony between the years 1843 and 1852. Of a rather more interesting character are Professor Karl Pearson's investigations into the results of Monte Carlo Roulette as recorded in *Le Monaco* in the course of eight weeks . . . on the hypothesis of the equi-probability of all the compartments throughout the investigation, he found that the actually recorded proportions of red and black were not unexpected, but that alternations and long runs were so much in excess that . . . *a priori* odds were at least a thousand millions to one against some of the recorded deviations. Professor Pearson concluded, therefore, that Monte Carlo Roulette is not objectively a game of chance in the sense that the tables on which it is played are absolutely devoid of bias. Here also, as in the case of Wolf's dice, the conclusion is solely relevant, not to the theory or philosophy of chance, but to the material shapes of the tools of the experiment.

\*

\*

\*

\*

We may now summarise the outcome of our enquiry at this stage in the following terms:

(i) The practical problem which gave rise to a formal calculus of probability is that of devising a rule for division of stakes in a game of chance with a view to ensuring a net gain in the long run to the gambler who adheres consistently thereto. The rule so conceived is a deduction from the premises assumed in the initial definitions of the concepts the calculus invokes. We can operate it only if we state it in advance, in which event we are *looking forwards*. In the initial formulation of the calculus we therefore have no sanction for *retrospective* judgments.

(ii) In the domain of games of chance, the Founding Fathers relied on their own intuitions to supply the missing link between numerical calculations derived from the algebra of choice and *observable frequencies* implicit in the assumed reliability of any rule conceived in the foregoing terms. The justification for their faith is amenable to empirical enquiry, if we circumscribe the terms of reference of the algebraic calculus in this way. If we invoke it to justify retrospective judgments, we raise a new issue which did not emerge in the context of the classical period, i.e. before Laplace propounded the doctrine of insufficient reason and therewith identified the algebraic concept with the subjective usage of the term probability in daily speech.

(iii) If we locate our definition of probability "in the mind" or define it without reference to external events, we have no firm foothold for asserting any class of factual situations to which the operations of the calculus are relevant. We may none the less formulate the operations of such a calculus with a conceivably useful outcome, if we embrace a responsibility which is the theme of Chapter 3, viz. that of investigating the property or properties common to the class of all situations in which it more or less correctly describes frequencies of observable occurrences in large-scale trials.

### CHAPTER THREE

## RANDOMNESS AND THE RELEVANCE OF THE RULE

IN THE FOREGOING CHAPTER we sought an answer to three questions:

- (a) To what class of problems did the Founding Fathers of the algebraic theory of probability conceive it to be relevant?
- (b) In what terms did they formulate the rules of the calculus and with what relevance to the end in view?
- (c) To what extent does experience of games of chance endorse their expectations?

We may summarise the outcome of our enquiry as follows:

- (i) the end in view is consistent with the terms of reference of the calculus, if, and only if, the definition of algebraic probability is factually identifiable with the long run *frequency* of the gambler's score;
- (ii) the concord of experience and theory vindicates the intuitions of the Founding Fathers that this is so in some situations, but not conspicuously in others;
- (iii) their explicit formulation of the theory fails to exhibit what characteristics of a situation endorse it as a more or less reliable code of conduct.

Our next task must therefore be to ask how to recognise such situations when we meet them. Here again we shall have reason to deplore how assuredly the clarification of statistical theory has been bogged down from the start by the emotive force of a vocabulary recruited from common speech. Otherwise, it would be difficult to explain the invocation of *bias* as an absolution for the failure of the theory to prescribe a reliable rule for division of stakes when experience does indeed discredit its claims; but the popularity of this verbal device does not belong to the classical milieu. Long after Laplace had given the algebraic theory a new orientation, writers on the algebraic

theory of probability accepted the factual credentials of the calculus at face value; and the experiments referred to in the preceding chapter belong to a period long after that of Laplace himself.

This hiatus between theory and practice provoked little concern till the mid-nineteenth century. Maybe, the invocation of bias to accommodate fact and fancy had already begun to provoke misgivings about the adequacy of the classical formulation; but the gambler had by then retired from the stage. While claiming new factual territories for occupation, writers on the mathematical theory of probability continued to rely on their intuitions about its relevance to the real world till Quetelet's programme (p. 18) of imperial stochastic expansion claimed the sanction of the empirical Law of the Constancy of Great Numbers, and thereby forced the inadequacies of the classical definition into the open. An intuitive formal approach in terms of the calculus of choice then made way for an ostensibly empirical definition of probability widely adopted after the publication of Venn's *Logic of Chance*, and more appropriate than the classical statement to the new use of the calculus by Maxwell. An earlier explicit formulation occurs in two contributions of Lewis Ellis, more especially "Remarks on the Fundamental Principles of the Theory of Probabilities" (*Trans. Camb. Phil. Soc.*, 1854). Keynes (*Treatise on Probability*) cites the last-named author's views as follows: "if the probability of an event be correctly determined the event will on a long series of trials tend to recur with frequency proportional to its probability."

If we are to approach the present crisis in statistical theory with a clear understanding of the relevance of the algebraic theory of probability to the long-term frequency of observable events, we must therefore digress from the historical sequence of its successive claims by seeking an answer to the question: *in what terms did mathematicians of the latter half of the nineteenth century seek to make the relevance of the rules more explicit by re-definition of the concept?* We here drop the idiom of the calculus of choice. Watching our gambler A who consistently sets his stake in accordance with a rule, we envisage a sequence of occasions (*trials*) at each of which an onlooker may make an

observation recording one or other outcome as an *event*  $E_1, E_2$ , etc. Venn speaks of the sequence as the *series of the event*. I shall speak of it more often as the framework of repetition. Among all trials of the sequence, we shall suppose that  $P_x$  is the proportion at which the onlooker correctly records the particular event  $E_x$ . We may suppose that A stakes to win in conformity with the rule: claim that the observed event is  $E_x$  at each trial. Then  $P_x$  will also be the proportion of truthful assertions the gambler will make if he consistently adheres to the rule regardless of the outcome of any run of trials in the sequence. In Venn's formulation  $P_x$  is the probability of the event. Chrystal, with a host of imitators, advances the following definition as substantially the view which Venn champions:

We are thus led to the following abstract definition of the Probability or Chance of an Event: If on taking any very large number ( $N$ ) out of a series of cases in which an event A is in question, A happens on  $pN$  occasions, the probability of the event A is said to be  $p$ . (*Text Book of Algebra*, Vol. II, 2nd edn., p. 567.)

In fairness to Venn and to Chrystal, one must add that *very large* in this context is a picturesque simplification of what is implicit in their usage of the definition. More precisely they signify by implication that  $p$  is the limit of a ratio  $s : N$  as  $N$  approaches  $\infty$ , if  $s$  is the actual number of occasions on which A occurs. Thereafter basic rules of addition and multiplication, etc., intrude ostensibly as corollaries. The numerous objections Keynes (pp. 104–6 *op. cit.*) puts forward and the particular criticism advanced by Aitken (*op. cit. infra*, p. 8) suffice to dispose of their claims as such.

Arguments commonly advanced against the Venn-Chrystal approach are formal; but we may dismiss it on other grounds. Since the intention of the concept is factual, it must embrace any actual situation consistent with the specification. Before we get to the corollaries, we find that Chrystal, as is true of any who follow Venn literally, has unobtrusively introduced factual qualifications and elaborations which are quite alien to his own definition as cited above, *inter alia* the following:

If for example, we assert regarding the tossing of a halfpenny, that out of a large number of trials heads will come up nearly as

often as tails—in other words, that the probability of heads is  $\frac{1}{2}$ , what we mean thereby is that all the causes that tend to bring up heads neutralise all the causes that tend to bring up tails. In every series of cases in question, the assumption, well or ill-justified, is made that the counter-balancing of causes takes place. That this is really the right point of view will be best brought home to us, if we reflect that undoubtedly a machine could be constructed which would infallibly toss a halfpenny so as always to land it head-up on a thickly sanded floor, provided the coin were always placed the same way into the machine; also, that the coin might have two heads or two tails; and so on. (*Op. cit.*, p. 568.)

The Gaussian gloss, i.e. the concept of causal neutralisation, superimposed on the Venn prescription in this context does not obviously guarantee that the behaviour of the penny conforms to the rules of a calculus of probabilities. It puts no restriction on the material composition of the halfpenny which might be of magnetised iron, copper-plated, to all appearances genuine and with faces of opposite polarity. Undoubtedly, we could then construct a machine to ensure that the penny would alternately fall heads and tails upwards. We have merely to suppose that the trials occur in a powerful magnetic field the polarity of which changes at each toss by photoelectric (or other appropriate) control. Clearly the ratio of heads to tails in this set-up approaches the 50 per cent limit more and more closely as the number of trials increases; and the gratuitously inserted concept of causal neutralisation holds good for a large enough sequence of trials, considered as a whole. Just as clearly, the rules of the calculus of probability, including the corollaries inconsequentially appended to the Venn-Chrystal formula, do not give a remotely correct description of the observable sequence of events, since the proportion of heads would be the same ( $\frac{1}{2}$ ) in every sequence of  $2r$  unit trials. To ensure that the end result is consistent with the corollaries, i.e. with the laws of addition and multiplication of probabilities, we have to postulate that the Gaussian principle of causal neutralisation operates at every *single* trial; but we have then added to our definition a new concept, that essential *lawlessness* which is characteristic of situations to which the so-called laws of probability are applicable.



From the classical viewpoint, the identification of probability with frequency at the level of definition in such terms is objectionable for another reason. We have undertaken to state in advance a rule which will be valid in an endless succession of trials, if the probability assigned to the event correctly specifies its limiting frequency in a sequence of games conceived as such. Now our definition identifies probability with a limiting frequency which it specifies on the implicit assumption that we either have at our disposal experience of an endless sequence, or may justifiably infer what will happen in such a sequence from other sources. Needless to say, our definition does not take such sources within its terms of reference.

In short, a definition of probability presupposing what Venn calls the *series of the event* must explicitly or implicitly contain something more than the postulate of a limiting ratio of observed occurrences of the series. Otherwise, and as Keynes rightly recognised, the derivation of the elementary rules of the calculus of probability entails a *non sequitur*. Contrariwise, the purely formal definition of probability in terms of what we now call the theory of sets is open to Venn's objection (*Logic of Chance*, p. 87) examined at length in the last chapter. Keynes (p. 94, *op. cit.*) states it more briefly thus:

When probability is divorced from direct reference to objects, as it substantially is by not being founded on experience, it simply resolves itself into the common algebraical doctrine of Permutations and Combinations.

We must now therefore come to grips with an issue which is more challenging than any we have so far faced. The question for which we must seek an answer is: *what initial assumptions must we endorse, if we wish to define the class of factual situations to which the rules of the calculus are indeed relevant?* In relying on intuitions consistent with long experience of the game, we must assume, for reasons already stated, that the Founding Fathers recognised that factual conditions such as the instruction to cut *anywhere* or to *shuffle* the pack are prerequisite to a fully explicit prescription of a betting rule for a card game. That such conditions also have implications not as yet disclosed will now emerge, if we re-examine what we do when we define the

probability of the event  $E_{x,r}$  referable to an  $r$ -fold sample as the ratio of all linear permutations of single events consistent with its specification to all  $r$ -fold linear permutations of single events in the infinite series of the event.

If we visualise how we can best set out all possible results of a 3-fold toss of a common die or of cutting our card pack (p. 47) of 6 spades (A, 2, 3, 4, 5, 6) three times, we may proceed as follows. First lay out, in each row of a square grid of 36 cells, dice or cards with 1, 2, 3 . . . 6 pips face upwards in ascending order from left to right, one to each cell. Next side by side with, and consistently left of or right of, one of the foregoing in each cell we also place in descending order in each cell of each column a die or card with 1, 2, 3 . . . 6 pips face upwards. We have now set out all  $6^2 = 36$  linear permutations of 6 distinguishable objects recorded 2 at a time without restriction on the repeated use of any one. In the same way, we may now lay out a new grid of  $36 \times 6$  cells, allocating to each cell of a row one of the 36 pairs and to each cell of a column one of the 6 cards or die-faces we may encounter at the third trial. We have then exhibited each of the  $6^3 = 216$  linear permutations of 6 distinguishable objects taken 3 at a time with replacement.

Evidently we can specify the 4-fold, 5-fold and in general  $r$ -fold choice by successive applications of this procedure. In following the iterative plan of the grid, we have then exhibited the two fundamental rules of the calculus implicit in the definition of p. 40. Thus there will be one way of scoring 18 in a 3-fold trial with probability  $\frac{1}{216} = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}$  in accordance with the theorem of multiplication. There will be 3 ways of scoring 2 twos and an ace each with probability  $\frac{1}{216}$ , and the probability of the event so defined is also by definition  $\frac{3}{216} = \frac{1}{216} + \frac{1}{216} + \frac{1}{216}$  in accordance with the rule of addition for exclusive events. There are also 3 ways of scoring two aces and a three. These 6 sequences exhaust the possibility of getting a total score of 5, whence by definition the probability of getting a score of 5 in the 3-fold toss is  $\frac{3}{216} + \frac{3}{216} = \frac{1}{36}$  in accordance with the addition rule.

Now what we do, when we thus exhibit the basic rules of the calculus of probability, i.e. the theorems of multiplication and

addition, as implicit in the classical definition of probability also discloses another implication we have not as yet examined. At each stage in the build-up of the  $r$ -fold sample we have conferred on every possible  $(r-1)$ -fold sample an equal opportunity of associating with each possible 1-fold sample. Regardless of the alternative methods of sampling we may postulate, specification of the correct number of different linear permutations consistent with a given sample structure thus signifies that *we allocate to each item an equal opportunity to associate in successive stages of the build-up of the sample with each residual item of the universe of choice*. We may speak of this assertion, which embodies the possibility of visualising the build-up of linear permutations by successive applications of a grid layout as the *principle of equipartition of associative opportunity*. In so far as a sequence of events in the real world guarantees such equipartition of associative opportunity, *if continued long enough*, we may speak of it as a *random* system; and in so far as it is a random system in this sense the probability assigned by the calculus to the event will prescribe its long-run frequency.

The relevance of our formal definitions to the sort of practical judgments with which we associate the word probability, then raises two questions:

- (i) whether there are any situations in which observation of a protracted sequence of events justifies the belief that the principle of equipartition adequately describes the limiting state of equilibrium;
- (ii) what characteristics of such situations, if they do indeed exist, are relevant to their recognition as such in real life?

When the Founding Fathers identified games of chance as one such class of situations, they had more than guesswork to go on. If the game involves a pack of cards, their intuitive conviction embraces the unstated assumption that we reshuffle the pack again and again lest previous disposition of the cards and inclination of the player to draw a card more or less near the top (or bottom) of the pack leads to unequal partition of opportunity for choice of one card rather than another. It is not obvious that this procedure will guarantee *randomness*, i.e. what

we have here referred to as equipartition of associative opportunity, if continued long enough; but it is at least a plausible assumption that it will do so. As it happens, experiments on card packs such as that of Pearson (p. 52) give the assumption a favourable verdict.

If we prefer to define the concepts of our calculus in formal terms unrelated to events, we may thus argue as follows: (a) experience vindicates the claim that probabilities assigned by the calculus closely tally with corresponding frequencies in long-term experience of an historical sequence of events in certain model situations; (b) in so far as we can impose on a system or predicate of it the relevant peculiarities of such model situations, we may justifiably extend its terms of reference to other situations. We then accept the obligation to impose randomness on, or at least to identify randomness as a property of, such situations. The distinction between imposing and identifying randomness will assume importance at a later stage, when we come to examine the concept of the infinite hypothetical population. Earlier remarks about the analogy between repetitive sampling from a card pack and the tossing of a die have anticipated it, but we must now accept the responsibility of scrutinising the analogy more closely.

In the domain of card packs or urn models we implicitly or explicitly recognise the need to take action, i.e. randomisation, on behalf of the calculus. In the domain of lotteries and die models, including the penny or the needle of Buffon, we are seemingly more reluctant to do so. Otherwise, it is difficult to explain the expenditure of so much industry (*vide infra*, p. 56) in a fruitless quest to vindicate the adequacy of the calculus to describe with equally conspicuous precision the long-run behaviour of an ordinary cubical die and that of Buffon's needle. The challenge evaded by blaming the die for its bias is not easy to rebut, if we retain the role of the passive spectator who accepts a toy in commercial production as a *fait accompli*; but we can carry our enquiry a step forward if we approach the issue in a more positive temper. We shall then gratefully dedicate ourselves to helping the calculus to work in return for so much work our contemporaries demand of it on our behalf.

At the outset, let us be clear about which of two questions

we are asking. To assert that the die behaves in a random way does not necessarily mean that it behaves in accordance with random sampling from a rectangular universe of six score classes; but it will at least be difficult to prove conclusively that it does so unless we can make this assumption with propriety. In any event, we shall require to specify the parent universe appropriately in advance, if the end in view is to prescribe a rule for the division of the stakes. We shall therefore interpret the challenge we have here accepted on the understanding that the specification of the six faces must indeed prescribe all the numerical information the classical definition incorporates. If no rule prescribed for the dealer then suffices to justify the use to which we put the definition, we may still entertain the possibility of enlisting a little co-operation from the manufacturer. Indeed, J. Bernoulli (*Ars Conjectandi*, Cap. IV) implicitly does so in disclosing the intuitive approach of his contemporaries:

. . . Evidently there are as many cases for each die as there are faces, and all these cases have an equal chance to materialize. For, by virtue of the similitude of faces and the *uniform distribution of weight* in a die, there is no reason why one face should show up more readily than another, as there would be if the faces had a different shape or if one part of a die were made of heavier material than another. (*Italics inserted.*)

We shall accordingly concede that there is an essential difference between Buffon's needle and the Chevalier's die. In either situation, we are observing the behaviour of a ponderable object heavier than air in a gravitational field. This of itself is entirely irrelevant to whether a needle will fall between or across the lines of Buffon's board. Unless the speed of the spin is exceedingly high, it is not irrelevant to the way in which a solid polyhedron will fall. The position of the centre of gravity will certainly influence the way in which it will most often come to rest unless located appropriately, and the construction of a homogeneous die with *indented* pips to make the article more durable certainly confronts the manufacturer with a very difficult task of location, if the end in view is to equalise the gravitational pull on the faces.

A trained investigator who set out to clarify the relevance of the calculus to a game of craps would not therefore design experiments of the sort summarised in Keynes's *Treatise*. *Inter alia*, he would carry out long-term trials to compare: (a) the results of spinning ordinary dice by machinery at very high speeds, letting ordinary dice roll gently to rest and tossing them negligently by hand in accordance with the gambler's practice; (b) the results of spinning in the ordinary way an ordinary die and a die with arbitrarily numbered faces distinguishable only by a thin homogeneous film of pigment of appropriate thickness and density;\* (c) the results when such dice and ordinary dice made to spin at the same speeds fall on smooth and on rough surfaces. In restating the problem of the die in this way, we have translated the risk of passively detecting properties which confer randomness on a system into the idiom of randomisation, conceived as an outcome of deliberate human interference. Our approach to a definition of randomness is then consistent with that of Peirce (*Theory of Probable Inference*) cited by Keynes (p. 290, *op. cit.*). A random sample is one "taken according to a precept or method which, being applied over and over again indefinitely, would in the long run result in the drawing of any one set of instances as often as any other set of the same number."

From a behaviourist viewpoint, the adequacy of our initial definition of probability or of our subsequent prescription of the relevance of the rules of the calculus in the real world is thus largely referable to how far and in what terms they make explicit both the concept of *randomness* and the situations of which it is a characteristic. Reluctance to get to grips with the issue in the formative stage of the mathematical theory of probability is perhaps explicable in its own setting. In the social context of Pascal and Fermat, probability was a calculus of the gaming table at which the nobility made or lost fortunes on the genteel understanding that the contest was truly a game of chance, and that intelligent specification of the correct wager would ensure eventual success. In conformity with the *mores* of

\* Alternatively, to make the article more durable each face might have six indents of equal capacity appropriately filled with white and black paint of equal density.

polite society and the dictates of good breeding, a gentleman did not probe too deeply into the unspeakable mystery of fair play; and a mathematician, if also a gentleman, could conveniently entrust the custody of the concept of randomness to his own conscience.

Such amiable conventions do not explain why some mathematicians of high repute and good judgment still decline the exacting invitation to clarify the semantic content of the word random by tinkering with the definition of probability in the hope that a supplementary and circular tautology will evacuate latent factual potentialities. Thus they permit us to define it in terms of a universe of choice only if it is *equally likely* that any residual item therein will occur in sample appropriately specified, or to define it in terms of a limiting frequency in the series of the event only if it is *equally likely* that any admissible subsequent event will follow any one specified antecedent. Textbooks still extant in the twenties of our century abounded in such emendations and evasions. Indeed Aitken (1939) finds it necessary to forewarn his readers against them, when he charitably states

Criticism is easy. The logician will not fail to pounce upon the words "equally likely," pointing out that they are synonymous with "equally probable" and that therefore probability is being defined by what is probable, a *circulus in definiendo*." (*Statistical Mathematics*, pp. 6-11.)

For expository convenience, Aitken himself prefers to confer axiomatic status on the concept of randomness. Concerning his own definition of probability he confesses

Something has been glossed over here; there is the tacit assumption that the initial phases are "equally likely". . . . The inclusion of the words . . . in a definition is in fact a concession: it *puts the reader more gently at terms with the abstract formulation* by anticipating its chief future application. The usage is not uncommon. . . . If a straight line is defined as "lying evenly" between its extreme points, what else does "evenly" mean but "in a straight line"? Every definition which is not pure abstraction must appeal somewhere to intuition or experience by using some such verbal counter . . . under the

stigma of seeming to commit a circle in definition." (*Italics inserted.*)

This is a candid and admissible viewpoint within its own terms of reference, i.e. class-room usage; but we are under no obligation to shirk the issue on that account. If we do indeed cherish the inclination to make the concept of randomness more explicit, it will be necessary to sidestep the snare of current idiom. When we use the expression *in the mind* the preposition unobtrusively identifies the dynamics of mental processes with mind conceived as a static object in three dimensional space. When we likewise speak elliptically of a *random sample* in contradistinction to *a sample taken at random*, we unwittingly predicate randomness as a property of an event in contradistinction to a property assignable to the series of the event. Within such a framework of discourse we must interpret the statement that events are *equally likely* to mean that they occur as often in the long run. We may then approach the terms of reference of the calculus from a different viewpoint.

We have seen that Venn's definition of the probability of an event in terms of frequency is inadequate because it fails to define what properties of the series of the event endorse the rules of the calculus. A definition which invokes the concept of frequency can indeed be adequate in that sense only if it takes within its scope an appropriate specification of the series itself. Accordingly, an algebraic theory consistently formulated in frequency terms must incorporate the concept of randomness in the initial definition of the series. Such is the comparatively recent restatement of the calculus by von Mises whose *Irregular Kollektiv* includes the notion of a limiting ratio in the empirical definition of Venn and adds the missing link in Venn's series of events, viz. the property we have here called the equipartition of associative opportunity. The concept embraces a sequence in which the number of distinguishable events may be finite or infinite. For illustrative purposes, it will suffice to conceive it in terms of 3 only:  $abbaacaaabccabbabbccacab \dots$ , etc.

If  $n_c$  here enumerates the event  $c$  in the first  $n_i$  events of the series, the ratio  $n_c : n_i$  becomes  $p_c$  when  $n_i$  becomes infinite in accordance with Venn's definition, but this of itself is consistent



with an orderly arrangement of the series. The new principle of disorder imposed on the series of the event by von Mises takes advantage of a conceptualisation of denumerable infinities not as yet widely current in Venn's time, viz. we can accommodate an infinity of even, an infinity of odd, an infinity of perfect squares, or an infinity of perfect cubes in one to one correspondence with the infinity of all integers, which specify the position (rank) of the event in the series. Thus one such infinite subsequence of  $n_5$  events is the subsequence of rank 5, 10, 15, 20 and so on. More generally, we say that there are  $n_{c,r}$  events specified as  $c$  in the finite subsequence of  $n_r$  events of rank  $r$ ,  $2r$ ,  $3r$  and so on. For any value of  $r$  we care to assign, the definition of von Mises specifies:  $n_{c,r} : n_r = p_c$ . Similarly, we might define  $n_5$  events of rank 1, 4, 9, 16, 25 and so on,  $n_p$  events of rank 1, 3, 5, 7, 11, 13, 17, 19, 23 and so on, and more generally  $n_0$  events selected in accordance with any orderly procedure. We still postulate  $n_{c,0} : n_0 = p_c$ .

Thus the proportionate contribution of the event  $c$  to the denumerable infinity of any prescribed subsequence is fixed. It is therefore entirely undetermined by its prescription; and this indeed we convey by the statement that the series of the event is *orderless*. The algebraic properties of the orderless series are formally identical with the operations of the calculus conceived in terms of set theory or of the classical concept of choice; but the initial definition of probability explicitly identifies it with frequency and also makes explicit the class of situations to which the rules are relevant, i.e. orderless systems of events. That it does indeed accommodate fundamental rules of multiplication and addition subsumed by the general term of the multinomial theorem in its alternative forms (p. 41) in conformity with the classical definition, will be evident from what follows.

We shall consider a sequence of  $n_i$  events, divisible into  $(n_i - 1)$  pairs of consecutive events. Within this sequence of  $n_i$  events, the particular event  $x$  occurs  $n_x$  times; and by definition  $p_x \simeq n_x \div n_i$  as  $n_i$  becomes indefinitely large, if  $p_x$  is the probability that  $x$  is the score at a unit trial. If  $x$  may denote any one of the score values  $b, c, d$ , we may write  $n_x = n_b + n_c + n_d$ . In the limit  $(n_b + n_c + n_d) \div n_i$  is then the probability that  $x$

will be one of the scores  $b, c, d$ . More explicitly we may then write  $p_x = P(b \text{ or } c \text{ or } d)$ , so that:

$$P(b \text{ or } c \text{ or } d) = \frac{n_b}{n_t} + \frac{n_c}{n_t} + \frac{n_d}{n_t} = p_b + p_c + p_d$$

This relation exhibits the additive property for unit trials, and we may extend it to compound trials by iteration. Thus we conceive  $n_{xyz}$  consecutive triplets (3-fold trials) out of  $(n_t-2)$  consecutive triplets of every sort in the  $n_t$ -fold sequence of single events. The probability that the triple event will be one of the score sequences  $abc, abd$  or  $bcd$  is the limiting value of  $(n_{abc} + n_{abd} + n_{bcd}) \div (n_t-2)$ . In the limit we may write  $(n_t-2) = n_t$ , and

$$P(abc \text{ or } abd \text{ or } bcd) = \frac{n_{abc}}{n_t} + \frac{n_{abd}}{n_t} + \frac{n_{bcd}}{n_t} = p_{abc} + p_{abd} + p_{bcd}$$

The derivation of the multiplicative property involves some circumlocution (*vide infra*), if we assume sampling without replacement from a finite universe. Otherwise, we may develop it iteratively from considerations of two consecutive events, of which  $(n_t-1)$  is the total number in the sequence of  $n_t$  single events. If  $n_{cx}$  is the number of pairs of which  $c$  is the initial event,  $n_{cx} = n_c$  unless  $c$  is the terminal event of the whole sequence, in which case  $n_{cx} = (n_t-1)$ . The probability ( $p_{cx}$ ) that the pair will have  $c$  as its initial member is by definition the limit of  $n_{cx} \div (n_t-1)$  and this is  $(n_c \div n_t) = p_c$ . The conditional probability ( $p_{b.c}$ ) that  $x = b$  is its successor if  $c$  is the initial event is also by definition the limit of  $(n_{cb} \div n_{cx}) = (n_{cb} \div n_c)$ ; and the unconditional probability ( $p_{cb}$ ) that a consecutive pair is  $c$  followed by  $b$  is the limit of  $n_{cb} \div (n_t-1)$ , which is  $n_{cb} \div n_t$ . We may thus write

$$p_{bc} = \frac{n_{cb}}{n_t} = \frac{n_c}{n_t} \cdot \frac{n_{cb}}{n_c} = p_c \cdot p_{b.c}$$

We may now reverse the order of procedure adopted in our last chapter to show that the general expressions cited on p. 41 follow from the additive and multiplicative properties of probability defined in the foregoing terms. We shall accordingly denote by  $p_{(s)}$  and  $p_s$  the probabilities of picking out a particular

score sequence  $S$  in exhaustive and repetitive sampling respectively; and may illustrate the build-up by reference to 3-fold sequences (AAB) in which two events of class A precede an event of class B. The multiplicative property then states that  $p_a \cdot p_{a.a} \cdot p_{b.aa} = p_{(s)}$  or  $p_s$  if we define  $p_{a.a}$  and  $p_{b.aa}$  appropriately. If sampling is without replacement from a finite universe of  $n$  items, of which  $a$  and  $b$  are respectively of classes A and B,

$$P_{(s)} = \frac{a}{n} \cdot \frac{a-1}{n-1} \cdot \frac{b}{n-2} = \frac{a^{(2)} \cdot b}{n^{(3)}}$$

If sampling involves no removal,  $p_{a.a} = p_a$  and  $p_{b.aa} = p_b$  so that

$$P_s = \frac{a^2 b}{n^3} = \left(\frac{a}{n}\right)^2 \left(\frac{b}{n}\right) = p_a^2 \cdot p_b$$

The probability that a sample will contain two items of class A and one of class B is that of getting any one of the sequences AAB, ABA, BAA for each of which  $p_{(s)}$  or  $p_s$  has the same value as above. By the addition theorem we therefore derive for the probability that the 3-fold score will be  $2A + B$  is

$$\text{without replacement } 3 \cdot \frac{a^{(2)} \cdot b}{n^{(3)}} ; \text{ with replacement } 3p_a^2 \cdot p_b$$

More generally, we may assume that  $n$  items consist of  $a$  of class A,  $b$  of class B,  $c$  of class C, etc., and that a sample of  $r$  items consists of  $u$  of class A,  $v$  of class B,  $w$  of class C, etc. For each sequence consistent with the specification, the theorem of multiplication means that

$$P_s = p_a^u \cdot p_b^v \cdot p_c^w \dots ; P_{(s)} = \frac{a^{(u)} \cdot b^{(v)} \cdot c^{(w)} \dots}{n^{(r)}}$$

There are  $r! \div (u! v! w! \dots)$  sequences consistent with each specification of  $u, v, w$ , etc. Whence we derive (iii) and (iv) of Chapter Two.

Such a restatement involves an expository difficulty which is not necessarily a disadvantage. In terms of the calculus of choice, the problem of sampling without replacement from a finite universe is on all fours with the problem of sampling with

replacement, and we may indeed exhibit the latter as a limiting case of the former when the universe becomes indefinitely large in comparison with the sample. The beginner who gets his first empirical ideas about probability from what happens in the allocation of hands of cards at whist or comparable situations, is therefore on familiar ground at the start. To conceptualise non-replacement sampling against the background of the irregular collective (I.K.) calls for some circumlocution which creates a difficulty for the beginner, but forces us to scrutinise the sampling process with greater circumspection. In effect, we regard each item taken from the static universe of choice as a lottery ticket which entitles us to identify the sampling procedure for a subsequent draw with a newly constructed I.K. My own staircase model of non-replacement sampling (*Chance and Choice*, Vol. I) is a visualisation of the combination of one I.K. with another in this way.

The tidiness of the formulation of a calculus of probability by von Mises is indisputably attractive; but the initial definition of the concept in terms of a frequency limit does not disclose any clue to an empirical criterion for its numerical evaluation by recourse to experience of a finite sequence of occurrences. This limitation, too easily overlooked, leads to verbal difficulties from which we can extricate ourselves only by the exercise of considerable circumlocution. If we embrace the classical or set theory definition on the understanding that we thereby undertake the obligation to provide experimental evidence of its relevance to the outcome of large-scale trials such as those cited in Chapter Two, we can readily attach a meaning to what Uspensky and other writers call the *constant probability of the event at each trial*, this being uniquely determined by the structure of a putatively fixed universe of choice. As defined by von Mises, the probability of the event is assignable as a property of the infinite collective of trials themselves; and it is difficult to compress the content of the words last cited in a compact form consistent with the definition of the concept.

As a basis for a betting rule for the benefit or comfort of the *individual* gambler, we have to conceive the orderless series of the event as a *temporal framework of repetition*, the specification of which is adequate to the task only if we have both: (a) some

reliable means of assigning in advance correct numerical values of the relevant basic parameters ( $p_a$ ,  $p_b$ , etc., in the foregoing exposition); (b) sufficient reason for assuming that the successive component unit scores of each game in an endless sequence of such games recur in an orderless succession. By accommodating the concept of randomness explicitly in his definition of probability, the algebraist has admittedly made explicit the class of situations with respect to which the calculus can claim to endorse a reliable code of conduct; but it is not the business of the algebraist to tell us how to recognise such situations when we meet them or whether we have at our disposal the requisite additional information prerequisite to a reliable rule for division of the stakes when we do so. Each of the conditions stated is essential. Thus we may satisfy ourselves that a particular system of successive shuffling ensures the prescribed disorder; but we cannot usefully apply this knowledge unless we also know the number of cards of each relevant class in the pack.

It is all too easy to lose sight of our equal obligation to equip ourselves with knowledge of both sorts when we interpret the *Kollektiv* in terms more relevant to anything we may legitimately or otherwise convey by assigning a probability to the truth of a scientific judgment. To speak of an individual field trial as one of an infinite succession of games is merely a figure of speech. Unless content to converse in metaphors, we must therefore leave behind us the temporal framework of repetition when we invoke stochastic theory in such situations. One way in which we can dispose of the need to interpret the series of the event as an historical sequence is to conceptualise the beneficiaries of the rule as an infinite team of gamblers each one simultaneously playing one game only with one of an infinitude of essentially identical packs, dice, etc., but the mere fact that the number of such dice, packs, etc., is accordingly limitless does not suffice to impose the condition of disorder on the system. The word *essentially* in this context will indeed be a danger signal, unless the reader has dismissed as trivial an earlier digression (pp. 44-47) on the analogy between the card pack and the die. As we there saw, our calculus will work if each player takes a sequence of  $r$  consecutive cards from

his assigned pack, only if we assume an adequate preliminary shuffle of each such pack.

Nor do we impose the condition of disorder on the conduct of the game, if we relinquish the right to discuss simultaneous sampling in a finite universe by conceding to each member of the infinite team of players the right to pick  $r$  cards simultaneously from one and the same pack containing a denumerable infinitude of cards suitably placed face downwards. We shall later see that this way of disposing of the inconvenience of conceptualising our *Kollektiv* exclusively in terms of an historic sequence has acquired the status of a widely current axiom, which doubtless derives a spurious cogency from the misinterpretation of an algebraic identity. Its correct interpretation in the domain of fact will not deceive us, if we are fully alert to the issues raised by the subversively over-simplified distinction between sampling with and without replacement.

We have seen that it is possible to subsume under two expressions (p. 41) the content of the classical definition of the probability of extracting an  $r$ -fold sample of specified constitution from a finite  $n$ -fold universe. Let us here recall them:

*without replacement*

$$P_{(u, v, w, \dots)} = \frac{r!}{u! v! w! \dots} \cdot \frac{a^{(u)} \cdot b^{(v)} \cdot c^{(w)} \dots}{n^{(r)}}$$

*with replacement*

$$P_{u, v, w, \dots} = \frac{r!}{u! v! w! \dots} \left(\frac{a}{n}\right)^u \left(\frac{b}{n}\right)^v \left(\frac{c}{n}\right)^w \dots$$

The formal identity of the two expressions if we substitute ordinary for factorial exponents or *vice versa* is an irresistible invitation to extensive generalisation. If  $n$  is very large in comparison with  $r$ , we may write  $a^{(u)} \div n^{(r)} \simeq a^u \div n^r$ , etc., and the two foregoing expressions become identical in the limit, i.e. for finite values of  $r$  when  $n$  is infinite. Actually, the prescribed probability of taking randomwise 5 specified cards simultaneously from a composite card pack of 200 ordinary full packs of 52 is assignable with a trivial error, if we specify it as that of recording the same 5 specified cards in a 5-fold cut

without removal; but the convenience of this approximation as a computing device does not dispose of a factual limitation fully discussed (pp. 43-49) in Chapter Two. If we cut 5 times randomwise, replacing the cards in their original order after each cut, the pack remains the same throughout the series of the event, and the formula holds good. If we pick a sequence of 5 cards simultaneously, we can do so randomwise only if we shuffle the pack before each game.

When our concern is with the 5-fold cut, the probability assigned to the compound event depends only on the proportions ( $p_a = a/n$  of class A,  $p_b = b/n$  of class B, etc.) of cards of relevant denomination in the pack. Thus the outcome of so-called replacement sampling is independent of the actual size of the pack; and the results of sampling in the infinite card pack without removal as commonly prescribed are the same as the results of such sampling in the 52-fold card pack, if the corresponding relevant parameters  $p_a$ ,  $p_b$ , etc., have identical values for each. By the same token, we may think of recording the toss of a hypothetical unbiased cubical die as comparable to cutting the pack of six cards on p. 47 or cutting an infinite pack containing cards with 1, 2, 3 . . . 6 pips in equal proportions. Having conceptualised our pack in this way, we are free to regard it as a widow's cruse from which we can continue indefinitely to extract finite samples without exhausting it, i.e. without changing the proportions of cards severally denoted by  $p_a$ ,  $p_b$ , etc.

Can we then say that taking a sequence of 5 cards simultaneously from any part of an infinite card pack is strictly comparable to cutting the pack anywhere five times? If we do so, we allow ourselves to be deluded by an algebraic trick. That it is indeed a trick becomes manifest, if we visualise how we actually choose the cards in the two situations. When we speak of cutting anywhere, the understood connotation of the adverb is that we cut in an orderless way. The act of cutting so prescribed thus imposes on the infinite series of the event the essential property of the *Kollektiv*, and we do not need to make any additional assumption about the arrangement of the cards in the pack, infinite or otherwise. As we have seen, we are free to regard the arrangement as fixed throughout, if we replace

the cards in the same order after cutting. That allowing  $n$  to become infinite does not restore this freedom, which we relinquish by prescribing a reshuffle of the  $n$ -fold card pack between each game defined as the extraction of an  $r$ -fold sequence of cards, is evident, if we specify one way of building up such a pack.

We have seen that we can choose blindfold with our two hands only 48 different 5-fold simultaneous sequences from one and the same full deck of 52 cards. If we now place on it a second full deck of cards arranged in the same order, we can actually choose 52, but we can never actually choose more than 52 different ones, if we pile on 3, 4, 5, etc., decks stacked in the same order. Our pack may become infinite, and we can still actually choose only 52 out of the  $52^5$  different sequences we could actually choose if free to shuffle the pack between each of a limitless sequence of games. On the assumption that the re-shuffle imposes the essential property of the *Kollektiv* on the act of choice, we should then legitimately deem the *replacement* formula to be appropriate to the situation. Otherwise, we must assume that each of the  $52^5$  different sequences occurs with equal frequency in the infinite stack.

To regard simultaneous sampling without removal as rightly equivalent to sampling with removal when the pack is infinite, does not therefore suffice to prescribe the way in which we choose the cards. We have also to introduce a new assumption about the arrangement of the cards in the pack. It is not enough to instruct the members of our infinite team to play the game in accordance with a rule of randomisation. We have also to equip them with an infinite card pack which has the unique property of randomness. As a model of sampling in field work, the infinite card pack does not, in short, exonerate us from the obligation to identify circumstances which endorse the putative relevance of the calculus of probability to the real world. Even if our statement of its initial terms of relevance incorporate the concept of randomness, we have still to settle the question: how can we prescribe randomness or recognise it when we meet it? So far, we have merely gleaned a clue to the first half of the question.

In so far as extensive experiments on games of chance have



subsequently vindicated the intuitions of the Founding Fathers, they do so only where human interference imposes the properties of randomness on a system; and it is at least difficult to see how such confirmation can endorse the intrusion of the calculus into the domain of retrospective judgment. Nor is it easy to derive from what we know about the manufacture of the die or of the lottery wheel, any sufficient empirical basis for detecting a system of complete disorder in nature. If we approach the factual credentials of the classical theory with due regard to the evidence assembled by Keynes, we shall not lightly assume that an organism or a society is comparable to a Prague lottery in 1754 or to a Brunn lottery in 1886 when what we also know about Monte Carlo is inconsistent with what we should prefer to believe about the habits of lotteries in general. Still less shall we complacently invoke the properties of such models to confer innate stochastic properties on natural phenomena for no better reason than our ignorance of what determines their vagaries.

The full statement of the views of von Mises first appeared in 1936. An English translation (*Probability, Statistics and Truth*) of the German text became available in 1939. With the introduction of the *Kollektiv* we thus arrive at the most recent attempt to make explicit the assumptions latent in the doctrine of the Founding Fathers with due regard to the end they had in view. Our enquiry has brought into focus the need to distinguish between what our contemporaries refer to as *randomisation* in the context of experimental design and *randomness* conceived as an innate characteristic of natural phenomena in the domain of statistical theory elsewhere referred to as a calculus of exploration. When we speak of randomisation we connote a property imposed by a plan of conduct on the collection of data, as when we prescribe the need to shuffle the card pack, shake the urn or spin the die. If we speak of randomness as the characteristic of a set-up in which we ourselves are passive spectators, nothing we have so far learned from classical models, i.e. games of chance, has given us a clue to its recognition.

If we do conceive randomness as an innate property of a die on all fours with its density or thermal conductivity, and seemingly von Mises does so, we have to admit that the concept

is factually meaningful only in so far as it is possible to prescribe how to construct a die whose long-term behaviour conforms to the requirements of the algebraic calculus. Experiment may well justify the belief that it is possible to do so; but the relevance of any such prescription to the recognition of randomness in nature is debatable. If it is possible to recognise randomness in nature, we must go to nature for our clues. That the outcome of doing so has not been wholly satisfactory to date sufficiently explains a widespread contemporary readiness to probe more deeply into the historical foundations of the algebraic theory. Meanwhile, the nostalgia with which our contemporaries turn to the writings of the Founding Fathers to reinforce their several claims as custodians of the classical tradition need not surprise us.

If our attempt to find in their own experience an unstated rationale for the intuitive convictions of Pascal, J. Bernoulli, de Moivre, Euler and their following has not been wholly successful, one reason may be that they would face an embarrassing dilemma, if asked to bestow their benediction on their successors of any persuasion. If called on to state the end in view, the answer they would have given would be unquestionably consistent with Neyman's interpretation of their intentions, that is to say the deduction of a rule which would ensure eventual success to the gambler if stated in advance and scrupulously obeyed thereafter. This intention does not encompass the right to evaluate a judgment referable to any isolated past event; and it is wholly unintelligible unless we conceive the formal definition of probability of an event in terms of the observable frequency of the event in a limitless number of trials. To that extent the outlook of the classical period is inconsistent with the repudiation of the notion of frequency by the school of Jeffreys and Carnap, and cannot accommodate the concept of fiducial probability (p. 441) embraced by the school of R. A. Fisher.

If asked to state explicitly why they believed that the formal identification of choice and chance is a sufficient basis for the deduction of any such rule, we can at best guess what answer the Founding Fathers would have given. No doubt, they would have conceded that each successively possible act of choice must

occur with equal frequency in the long run; but if forced to justify this faith one may well suspect that they would have been able to enter no better plea in defence than our ignorance of the outcome. The gratuitous assumption that events which are equally likely are events about which we know nothing is the penalty we pay for imposing a numerical specification in a technical context on a word which signalises by daily and long association an unquantifiable measure of uncertainty. To that extent, the Founding Fathers might well have felt themselves on familiar ground in discussing the fundamentals of the theory in the idiom of Jeffreys and Carnap. Though he consistently restricts his theme to an examination of the observable long run frequency of external events, Bernoulli, in whose *Ars Conjectandi* the classical theory as set forth in the first half of the preceding paragraph first takes shape in its entirety, equates in several places the conviction that a die will behave in a particular way to the assertion that we have no reason for believing the contrary.

In the restatement of the theory by von Mises this axiom drops out; and we must concede that his definition of probability, unlike any previously put forward by a professional mathematician, formally specifies the class of situations to which the calculus is truly relevant. Unluckily, the specification is not sufficiently explicit at the factual level to disclose how to set about identifying such situations when we meet them. Nor is it easy to give a confident prescription for doing so by relying on what is factually common to situations in which the calculus does work as well as we have reason to hope. Any such situations so far examined involve an agent and a repetitive programme of action. If, as is true of card pack and urn model situations, the trial involves an act of choice in the most literal sense of the term, the instructions embodied in the programme must include an explicit specification of the act itself; but in any case they will embrace instructions for shuffling, tossing or spinning with the ostensible aim of imposing disorder, as von Mises uses the term, on what Venn calls the series of the event. The execution of such instructions constitutes the putative randomising process. The identification of such a programme of instruction with the concept of randomisation forces us

therefore to carry our enquiry a step further by asking: what is common to all such programmes?

At this stage, we can plausibly predicate only three features shared by all of them. All impose the condition that the sensory discrimination of the agent is impotent to influence the outcome of the programme. All embrace the possibility that strict adherence to the programme is consistent with the realisable occurrence of any conceptually possible outcome in a single trial. All exclude the possibility that any orderly rhythm of external agencies intervene to impose a periodicity on the series of the event. Whether these three features suffice to specify a randomising process as such is difficult, if not impossible, to prove; but the indisputable relevance of the first to a satisfactory specification confers no sanction on the principle of insufficient reason. The latter endows the subjective judgment of the passive spectator with the property of generating randomness without active participation in promoting the series of the event.

## CHAPTER FOUR

### DIVISION OF THE STAKES AND THE LOTTERY OF LIFE AND DEATH

THROUGHOUT THE CLASSICAL PERIOD, i.e. from the correspondence of Pascal with Fermat to the enunciation of the doctrine of inverse probability by Laplace, the hazards of the gaming table were the major preoccupation of mathematicians who contributed to the formulation of a stochastic calculus. Whatever views we may entertain about the relevance of the calculus of probability to scientific enquiry and in whatsoever way we may now choose to define probability in their context, there can thus be no doubt about the intentions of the Founding Fathers. Their practical aim was to prescribe a betting *rule* to ensure a profit to the gambler who consistently applies it. A rule so conceived is a rule the gambler must continue to apply in spite of a run of bad luck. It therefore dictates inexorably the proper form of the assertion which specifies the gambler's bet, even if the fortunes of the gambler himself tempt him to doubt its reliability. Only on that understanding can we assign a probability to the truth of his assertion. Clearly, therefore, the proper use of the theory conceived in such terms cannot accommodate the right to assign a probability to an assertion ostensibly justified by the occurrence of a particular event in the past; but it would not be true to state that all of the predecessors of Laplace invariably recognised this limitation of the classical formula. In a prize essay of the *Academie Royale des Sciences*, D. Bernoulli (1734) discusses at length whether the narrow zone in which the orbits of the planets lie around the ecliptic is attributable (in Todhunter's words) to hazard. If we sustain the Forward Look, this is a meaningless query. The planets lie where they do lie.

Such lapses are exceptional. Nor did they fail to provoke intelligent protest at the time. Even in the treatment of a new theme which intrudes as we approach the grand climacteric of the classical theory, the end in view is seemingly consistent with the programme stated in the preceding paragraph; and

if the justification for extending the terms of reference of the theory to embrace it is exceptionable, objections which now seem self-evident invoke factual considerations far less familiar at the time. This new theme which now invites our scrutiny is a class of problems which to writers of the century antecedent to the promulgation of his eternal law of population by parson Malthus might well seem to be cognate to the hazards of the gambler. In the sixteenth century the practice of insurance had indeed grown gradually out of a traffic in wagers at mediaeval fairs. It expanded briskly in the eighteenth century against the background of Halley's Life Table published in the *Philosophical Transactions of the Royal Society* (1693). As a Huguenot refugee after the Revocation of the Edict of Nantes, de Moivre undertook actuarial work to support himself, and annuity expectations make their appearance in his *Doctrine of Chances* (1718) in the same milieu as the fortunes of the gaming saloon. The treatise itself deals largely with the latter; but the identification of insurance risks with the hazards of gambling asserts itself more obtrusively in later writings of the classical period, including in particular the works of Euler and of d'Alembert.

In what sense, if any, the classical theory of the gambler's risk offers a rationale for the costing of an insurance corporation will be the more easy to recognise, if we examine the implications of a theorem announced by J. Bernoulli (1713) in the *Ars Conjectandi*. To set forth the theorem in formal terms, we may specify the probability that the sample mean score ( $M_0$ ) referable to a distribution whose true mean is  $M$  lies in the range  $M \pm \epsilon$  as:  $P(M - \epsilon \leq M_0 \leq M + \epsilon) = (1 - \alpha)$ . Then  $\alpha$  is the probability that a deviation is numerically greater than  $\epsilon$ ; and the value assignable to  $\alpha$  depends both on  $\epsilon$  and on the size ( $r$ ) of the sample. For a fixed value of  $r$ , deviations numerically greater than  $\epsilon = u$  will be more frequent than deviations numerically greater than  $\epsilon = v$  if  $v > u$ ; and we must assign a lower value to  $\alpha$  if we assign a higher value to  $\epsilon$ . For one and the same value of  $\epsilon$  we may assign a lower value to  $\alpha$  if we assign a higher value to  $r$ ; and the probability that  $M_0$  will lie inside a range specified as above will approach unity as  $r$  approaches infinity. The theorem states that this is true for any value of  $\epsilon$  *however small*.

So stated, it admits of a more or less general proof by several methods; but it will here be convenient to approach it in a way which facilitates numerical calculations illustrative of its bearing on one factual preoccupation of its author. We shall then see more clearly what it does *not* mean. Accordingly, we shall take advantage of the fact that the normal curve provides a close fit for the random sampling distribution of the mean score of the  $r$ -fold sample from any  $n$ -fold discrete universe of choice, if  $r$  is very large but also a small fraction of  $n$  (Appendix I). On this understanding, we shall denote by  $\sigma_m^2$  the variance of the  $r$ -fold sample mean score distribution and by  $\sigma^2 = r\sigma_m^2$  the fixed variance of the unit score distribution. We then define a normally distributed standard score of unit variance by  $h = (M_0 - M) \div \sigma_m$ , whence if  $h\sigma_m = \epsilon = (M_0 - M)$ , we may write:

$$\begin{aligned} P(M - \epsilon \leq M_0 \leq M + \epsilon) &= (1 - \alpha) \\ &= P(M - h \cdot \sigma_m \leq M_0 \leq M + h \cdot \sigma_m) \quad (i) \end{aligned}$$

For the probability that  $M_0$  lies within the limits  $(M \pm h\sigma_m)$  when  $h$  is positive, we may write:

$$\begin{aligned} P(M - h \cdot \sigma_m \leq M_0 \leq M + h \cdot \sigma_m) \\ = (1 - \alpha) = (2\pi)^{-\frac{1}{2}} \int_{-h}^h e^{-\frac{1}{2}c^2} \cdot dc \end{aligned}$$

$$P(M_0 < M - h\sigma_m) = \frac{1}{2}\alpha = P(M_0 > M + h\sigma_m)$$

$$P(M_0 > M - h\sigma_m) = (1 - \frac{1}{2}\alpha) = P(M_0 < M + h\sigma_m)$$

For illustrative use in what follows, we may obtain from the table of the normal integral the following values for  $\alpha$  in terms of  $h$ :

$h$	$(1 - \alpha)$	$\frac{1}{2}\alpha$	$(1 - \frac{1}{2}\alpha)$
1.64	0.8990	0.0505	0.9495
1.96	0.950	0.0250	0.9750
2.43	0.985	0.0075	0.9925
2.81	0.995	0.0025	0.9975
3.08	0.998	0.0010	0.9990
3.20	0.9986	0.0007	0.9994

In (i) above we have defined  $\epsilon$  in terms of  $h$  and  $\sigma_m$ , but we

may define it alternatively in terms of the fixed constant  $\sigma$  and  $r$ , viz.:

$$\epsilon = \frac{h\sigma}{\sqrt{r}} \quad \text{and} \quad r = \frac{h^2\sigma^2}{\epsilon^2}$$

The formal meaning of the theorem is now clear. We can fix  $\alpha$  in (i) at any level, however small, by making  $h$  sufficiently large, e.g.  $\alpha = 0.005$  if  $h = 2.81$ ,  $\alpha = 0.002$  if  $h = 3.08$  and if  $h = 3.2$ ,  $\alpha \simeq 0.0014$ , as shown in the foregoing table. Having thus fixed  $\alpha$ , we may then make  $\epsilon$  as small as we like by increasing  $r$  sufficiently. On the assumption that the calculus is in fact consistent with the natural history of the game, we can thus ensure that the frequency of a deviation of the sample mean score ( $M_0$ ) from the true mean ( $M$ ) numerically greater than any minute quantity  $\epsilon$  however small will itself be inappreciably small, if we make the size of the sample, i.e. the number of unit trials per game, sufficiently large.

The statement last made emphatically does not mean that enlarging the size of a finite sample will ensure that deviations numerically larger than  $\epsilon$  will *never* occur. Thus it is difficult to see why so many writers exhibit results such as those set forth on pp. 52-57 above as empirical verification of Bernoulli's theorem. We shall prefer to regard them as positive evidence for the belief that the operations of the stochastic calculus do creditably conform to the long-run luck of the game; and to that extent they are justification for using the theorem as a factually legitimate extension of its terms of reference. As Keynes remarks, the theorem in its own right is amenable neither to proof nor to disproof, the record of a large deviation in a single large-scale experiment being consistent with what the theorem states may occasionally, albeit very rarely, happen.

What the theorem rightly signifies in the social context of the *Ars Conjectandi* will become more clear if we examine a model situation. We suppose that a lottery wheel:

- (a) has 10 sectors of equal area, nine black and one red;
- (b) each sector comes to rest against a pointer with equal frequency in the long run;
- (c) a gambler A scores 1 when the red sector comes to rest against it and zero otherwise.



In this set-up the true mean score of the distribution is  $M = 0.1$  and  $\sigma = \sqrt{(0.1)(0.9)} = 0.3$  for the unit trial distribution. Successive terms of the binomial  $(0.9 + 0.1)^r$  exactly specify the sampling distribution of the mean score ( $M_0$ ) of the  $r$ -fold spin with variance  $\sigma_m^2 = (0.3)^2 \cdot r^{-1}$ . If  $r > 300$ , the normal quadrature will be very close and we may then regard  $h = (M_0 - M) \div \sigma_m$  as a normal score of unit variance. By definition therefore

$$(M_0 - M) = \epsilon ; \epsilon = h\sigma_m = -\frac{3h}{10\sqrt{r}}$$

$$\therefore r = \frac{9h^2}{100\epsilon^2} \text{ and } h = \frac{10\epsilon\sqrt{r}}{3} \quad (\text{ii})$$

Thus we may first suppose that  $\epsilon = 0.03$ , so that  $\alpha$  is the probability that the player's mean score lies outside the range  $0.07 \leq M_0 \leq 0.13$  and  $r = 100h^2$ . From the foregoing figures citing  $\alpha$  in terms of  $h$ , we then derive

$h$	$\alpha$	$r$
1.96	0.05	384
2.81	0.005	790
3.08	0.002	949
3.2	0.0014	1,024

The theorem acquires a new interest if we use (i) and (ii) above to formalise a rule for division of the stakes, invoking the notion of the gambler's *expectation* of gain and the *risk* associated with a specified gain or loss. We suppose that A agrees at each spin to pay \$ $y$  forfeit if the score is zero, receiving \$ $x$  compensation if his bet is right, i.e. if the score is unity. Thus he either gains \$ $x$  or loses \$ $y$  at each spin. If  $M_0$  is the proportion of successes in a game of  $r$  spins, the net gain is

$$G = r \cdot M_0 \cdot x - r(1 - M_0)y$$

For brevity we may write  $M = p = (1 - q)$  and  $\epsilon = (M_0 - M)$

$$\therefore G = r(px - qy) + r\epsilon(x + y) \quad (\text{iii})$$

The theorem implies that  $\epsilon$  approaches zero with a probability ever nearer unity as  $r$  approaches infinity. In the limit therefore

$$G_r \simeq r(px - qy)$$

We speak of  $E = (px - qy)$  as the gambler's expectation of gain per game, and may write (iii) as

$$G = r [E + \epsilon(x + y)] \quad (\text{iv})$$

When  $\epsilon = 0$  so that  $M_0 = M$ , the total gain is  $rE$  and

$$P(G < rE) = 0.5 = P(M_0 < M) \quad (\text{v})$$

Now the gambler must lose in the long run, if the expectation is negative, i.e. if  $E < 0$  so that  $px < qy$ . If his opponent agrees to pay 10 dollars for each success A scores, A will thus win in the long run if  $10p > qy$  and lose if  $10p < qy$ . In the set-up referred to above,  $p = 0.1$  and  $q = 0.9$ . Thus the forfeit A will agree to pay must be  $y < 1.1$ , if he hopes to win in the long run. If  $y = 1.1$ , so that  $G = 0$  for  $\epsilon = 0$ , the probability of eventually losing the game is therefore 0.5; but if A sets his stake with his opponent's consent at 1 dollar ( $y = 1$ ), so that  $E = 0.1$ , he has a 50 per cent chance of making at least  $(0.1)r$  dollars in a game of  $r$  spins. Nor will he necessarily lose in the long run if  $M_0 < M$ , so that  $\epsilon$  in (iv) is negative. So long as  $E$  is numerically greater than  $\epsilon(x + y)$ , his net gain will be positive. If we write  $\epsilon = -k = h\sigma_m$  the risk of losing the game is then

$$P(M_0 < M - k) = \frac{1}{2}\alpha = P(M_0 < M - h\sigma_m)$$

When  $\frac{1}{2}\alpha = 0.001$ ,  $h = 3.08$ , so that

$$k = \frac{3(3.08)}{10\sqrt{r}} \quad \text{and} \quad r = \frac{0.853776}{\epsilon^2}$$

If A wants to keep the risk of losing the game at the 0.001 level for the stakes,  $x = 10$ ,  $y = 1$ , in which event  $k = 0.009 \simeq (110)^{-1}$ , the duration of the game (number of spins) will be

$$r = (110)^2(0.853776) \simeq 10,331$$

We shall now suppose that the rule fixes the duration of the game as  $r = 625$ , so that

$$k = \frac{3(3.08)}{250} = 0.03696 \quad \text{and} \quad 11k \simeq 0.4065$$

Since  $(x + y) = 11$  for the stakes  $x = 10$  and  $y = 1$ , his net gain for  $11k = 0.4065$  is

$$1(0.1 - 0.4065) = -625(0.3065) = -192$$

Thus  $0.001$  is the risk that he will lose at least 192 dollars if the game lasts for only 625 spins. To keep the risk of loss at the same level he must therefore get his opponent to agree to an offer of a lower stake  $y$  for the same stake  $x = 10$ . Accordingly, we may write:

$$E = 1 - (0.9)y \quad \text{and} \quad (x + y) = (10 + y)$$

Hence A neither gains nor loses if  $E = -\epsilon(10 + y)$  and  $\epsilon = -k$  is negative, so that

$$k = \frac{1 - (0.9)y}{10 + y}$$

The probability that he will not win is

$$P(E \leq 10k + ky) = \frac{1}{2}\alpha = P(M_0 \leq M - k)$$

If  $k = h\sigma_m$  and  $r = 625$ ,  $k = 3h \div 250$  and  $h = 3.08$  if  $\frac{1}{2}\alpha = 0.001$ , so that  $k = 0.03696$  if the risk is  $0.001$

$$\therefore (0.03696)(10 + y) = 1 - (0.9)y$$

$$\therefore y \simeq 0.673$$

So far we have assumed that he fixes the stake in accordance with a pre-assigned risk of not winning the game. We may also fix the stake at a pre-assigned risk on the assumption that he will gain at least  $m$  dollars or lose less than  $d$  dollars. If he wishes the risk to be  $0.001$  that he will not make at least 250 dollars, we may illustratively\* proceed to evaluate  $y$  for  $x = 10$  as follows. To win 250 dollars we must assume:

$$625 [1 - (0.9)y - k(10 + y)] = 250$$

$$P(G < 250) = 0.001 = P(M_0 < M - k)$$

\* At the level  $h = 3$  for  $r = 625$  the normal quadrature is by no means reliable when  $p = 0.01$ .

If  $\frac{1}{2}\alpha = 0.001$ ,  $h = 3.08$  as before and  $k = 0.03696$ , so that

$$\begin{aligned} 625 [0.6304 - 0.937 \cdot y] &= 250 \\ \therefore y &\simeq 0.246 \end{aligned}$$

Thus his stake will be roughly a quarter dollar.

We may thus summarise as follows the general theory of the division of the stakes, when we may assume that the normal quadrature is sufficient. If  $G$  is the gain of gambler A in a game of  $r$  unit trials,  $x$  the stake his opponent B pays up if A wins his bet,  $y$  the forfeit A pays up if he loses it,  $p = M = (1 - q)$  the probability that A will win in a single trial:

$$G = rE + r \cdot \epsilon (x + y) \quad \text{in which} \quad E = px - qy$$

In this expression  $p$  and  $q$  are constants assigned by the nature of the game, and we may also assume that the opponent fixes the stake  $x$  in advance. We may then consider  $r$ ,  $y$  and  $\epsilon$  as unknown. We may specify the risk that A will not win as much as a fixed sum  $m$  in the form:

$$P(G < m) = \frac{1}{2}\alpha = P(M_0 < M - k) \quad (\text{vi})$$

The relation between  $k$  and  $m$  is then

$$m = rE - rk(x + y) \quad (\text{vii})$$

If we write  $k = h\sigma_m$ , we may evaluate  $\frac{1}{2}\alpha$  by recourse to the identity  $\sigma_m^2 = r^{-1}p(1 - p)$  which fixes  $h$ . Conversely we may fix  $\frac{1}{2}\alpha$  at any assigned level by first fixing  $h$ . Thus the probability of not winning as much as  $m$  is assignable as a known function of  $h$ ,  $r$  and  $y$ , i.e.

$$P(G < m) = F(h, r, y)$$

If we wish to assign the risk ( $\frac{1}{2}\alpha$ ) for fixed stakes ( $x, y$ ) and fixed duration ( $r$ ) of game for less than a fixed gain  $m$ , we solve (vii) for  $k$  and hence for  $h$ , having pre-assigned  $x, y$  and  $r$ . If we wish to specify the stake  $y$  for a fixed risk in a game of fixed duration, we pre-assign  $h, x, r$  and likewise appropriately for  $\frac{1}{2}\alpha$ , whence also  $k$ , to solve (vii) for  $y$ . If we wish to determine the appropriate duration of the game for fixed stakes at a

pre-assigned risk, we solve (vii) for  $r$  having pre-assigned  $x$ ,  $y$  and  $h$ , whence  $k$ .

The rule for the division of the stakes assumes a new interest if we introduce a third gambler to underwrite A. We have seen that A's risk of losing the game of 625 spins is about 0.001, if he stakes 67.3 cents against 10 dollars. In that event he cannot lose more than  $625(0.673) = 421$  dollars per game. In  $n$  games he stands to lose at most  $421 n$  dollars and to gain as much as  $6,250 n$ . We may suppose that his capital is 105,250 dollars. He will therefore remain solvent, if he plays no more than 250 games with a possibility of winning as much as 1,562,500 and of thereby increasing his capital to 1,667,750 dollars. We may also suppose that a third gambler C with more capital and many such clients agrees to pay any debts A may incur, if A pays him 5 dollars a game. Thus A may lose  $(421 + 5)$  dollars per game or  $426n$  in  $n$  games. He must then remain solvent if he plays less than 248 games with the possibility of increasing his capital by 1,542,515 dollars. In this set-up, C may have to pay up as much as  $(247)(416) = 102,752$  dollars and may add as much as  $5(247) = 1,235$  dollars to his capital.

Let us now regard C as a corporation with enough capital to take on 10,000 clients on the same terms as A without the possibility of insolvency. In that event, we may put its gain per successful game as  $x = 5$  and its loss at a figure never exceeding  $y = 416$ . If the stake were exactly 416, we might write the expectation accordingly as:

$$E = 5(0.999) - 0.001(416) = 4.579$$

On the same assumption, there are therefore equal odds that C will gain more or less than 45,790 dollars; but the figure 416 is not the actual stake which C forfeits for a failure, being the greatest forfeit C can ever pay in a game with a variable forfeit ( $y$ ) for a fixed return ( $x$ ). Even so, the risk that C will not substantially gain is small.

For illustrative purposes, we shall thus put the plight of C *at its worst*, if we assume that C always pays up 416 dollars when A loses. We shall then think of C as standing to lose 416 dollars with probability 0.001 and to gain 5 dollars with probability 0.999 in each of 10,000 unit trials. The net gain

in 10,000 games (one per client) is accordingly specifiable as:

$$45,790 + 10,000\epsilon(421)$$

For  $p = 0.001$  and  $r = 10,000$ ,  $\sigma_m \simeq 0.0003162$ .

If  $h = 1.64$  so that  $\frac{1}{2}\alpha = 0.05$ , we set 5 per cent as the risk that C will gain less than

$$45,790 - 10,000(421)(0.0005186) \simeq 43,607$$

Needless to say, this is a gross underestimate of what C stands to gain with 5 per cent risk of failure. In any event, the possible loss the corporation might incur is  $(102,752) \times (10,000) = 1,027,520,000$  dollars, and without danger of insolvency if this is indeed its capital. Having less capital, it may insure itself against insolvency by restricting the number of its clients and paying 50 cents per client to a more wealthy corporation D to cover all possible loss. It can then take on  $c$  clients with no possibility of insolvency if

$$(0.5)c + (102,752)c = 1,027,520,000$$

Thus C may still accept nearly 10,000 clients with complete protection against insolvency if it pays to D a premium of 10 per cent on all premiums received from gamblers who operate in the same way as A.

The foregoing is more than a parable. During the latter half of the eighteenth century the government of Royalist France reaped a handsome revenue from a lottery which throws light on the predilection of Laplace for the urn model. Other European governments operated similar state lotteries in the following century. A French citizen of the earlier period could purchase one or more *billets* numbered from 1 to 90 inclusive. At appointed intervals an official drew randomwise 5 tickets of a complete set of 90. On the announcement of the result the holder of one or more tickets bearing the same number as a ticket drawn could claim compensation. Uspensky tells us that the holder of one or more tickets with the same winning number could claim 15 times the cost of each, of one or more pairs with 2 different winning numbers 270 times the cost of each, and so on. For each claim so specified, the government set its

stake to gain in the long run. If  $N$  is the number on a single ticket, the probability that one of 5 taken from 90 will be  $N$  is  $\frac{5}{90} = p = \frac{1}{18}$ . If the holder pays  $t$  francs for it, the government thus gains  $t$  with probability  $\frac{17}{18}$  if  $N$  is not one of the winning numbers and otherwise loses  $(15 - 1)t$  with probability  $q = \frac{1}{18}$ . The expectation of the government is therefore positive, being

$$E = \frac{17}{18}t - \frac{14}{18}t = \frac{1}{6}t$$

If the holder has two tickets with different numbers  $N$  and  $M$ , the probability that the second is a winner if the first is also is the probability that it is one of 4 out of 89 tickets and  $p = \frac{5 \times 4}{90 \times 89} = \frac{2}{801}$ , and this is the probability that the government will have to pay  $2t(270)$  losing  $2t(270) - 2t = 538t$ , whence

$$E = \frac{(799)2t}{801} - \frac{2 \cdot 538t}{801} = \frac{58}{89}t$$

If the government sold single tickets to 100,000 different holders at 10 francs, its net gain in francs in accordance with (iii) above would be

$$G = 1,000,000 \left( \frac{1}{6} + 15\epsilon \right)$$

For this set-up  $p = \frac{17}{18}$  and

$$\sigma_m = \frac{\sqrt{17}}{1800\sqrt{10}} \simeq 0.000724$$

Whence  $\epsilon \simeq -0.002231 = h\sigma_m$  if  $h = 3.08$  and we put  $\frac{1}{2}\alpha = 0.001$  as the risk of making less than

$$1,000,000 \left( \frac{1}{6} - 0.0335 \right) = 133,200 \text{ francs}$$

The risk of making no profit is  $\frac{1}{2}\alpha = P\left(15k > \frac{1}{6}\right)$  and  $15k > \frac{1}{6}$  if  $h > 15.3$ , whence  $\frac{1}{2}\alpha < 0.000000001$ .

Till abandoned in 1789, the enterprise proved profitable to the regal gambler. Had the monarch chosen to limit the issue of tickets and to seek coverage against insolvency like Gambler A in the foregoing parable, we may suppose that an insurance company C with more capital than the treasury might also have profited by underwriting possible loss and have covered itself against insolvency by a comparable arrangement with one of the big banking houses (D) in a period when a banking family could (and did) in effect underwrite the risk that Wellington would lose the Battle of Waterloo. Such then was the setting in which one of the first surviving Life Insurance corporations, the *Equitable*, came into being (1762).

There had been earlier and unsuccessful ventures. If they had conducted business solely with lottery proprietors, Bernoulli's theorem might supply a rationale for success of one or failure of the other. A gambler with more capital can play more games without risk of insolvency and play them for higher stakes. More often then, but not necessarily so, the firm with more capital will increase its capital and less often will become insolvent than a smaller one, meanwhile improving its prospect of continuing in business with greater rewards, if it does not fail. This interpretation of the theorem presumes that the firm does in fact operate in circumstances strictly comparable to the relation between the corporation C and the gambler A. That gambling on the lives of one's clients is strictly comparable to gambling on the success of a lottery or of a gamester at the card table will seem plausible enough if we equate what is probable to what is not quite certain, whence and by merely verbal association extending the terms of reference of the algebraic concept of probability to embrace any situations about which our information is inadequate to sustain a firm assertion. The author of the theorem (Cap IV) does indeed anticipate such a use of it when he thus foreshadows a stochastic rationale of the Life Table:

. . . who can say how much more easily one disease than another—plague than dropsy, dropsy than fever—can kill a man, to enable us to make conjectures about the future state of life or death? . . . such and similar things depend upon completely hidden causes, which, besides, by reason of the innumerable variety of combinations will



forever escape our efforts to detect them. . . . However, there is another way to obtain what we want. And what is impossible to get a priori, at least can be found a posteriori; that is, by registering the results of observations performed a great many times.

To assess rightly the relevance of the theorem to the prospects of the *Equitable* corporation, we may usefully recall Uspensky's verdict, and his two criteria of its legitimate application

. . . little, if any, value can be attached to practical applications of Bernoulli's theorem, unless the conditions presupposed in this theorem are at least approximately fulfilled: independence of trials and constant probability of an event for every trial. And in questions of application it is not easy to be sure whether one is entitled to make use of Bernoulli's theorem; consequently, it is too often used illegitimately.

That year to year variation of death rates referable to particular diseases is strictly independent in the stochastic sense of the term is a postulate to which biologists in general and epidemiologists in particular could now advance formidable objections. To assert that the probability of the event is fixed from year to year is flagrantly inconsistent with mortality experience of all civilised communities in the century and a half dating from the eclipse of the classical tradition, but such an assumption was by no means repugnant to common experience in the historic context of the *Ars Conjectandi*. If therefore the historical tie-up between insurance and gambling helps us to understand how easily this identification could gain universal assent, a necessary logical relation between the classical calculus of probability and the empirical rule of thumb which dictates actuarial practice is today less easy to detect.

In our own social context, it is surely difficult to see why computations referable to empirical confidence in the short-term stability of the vital statistics of populations requires any explicit introduction of the classical theory. The entire argument on which Euler (1760) bases his formulae for life annuities in *Récherches sur la mortalité*, etc., is presentable, like those for the construction of the life table itself, without recourse to any considerations other than experience and simple arithmetic.

Indeed, this is how Todhunter expounds it. Todhunter himself expresses an unwitting uneasiness when he declares

This history of the investigations on the laws of mortality and of the calculations of life insurances is sufficiently important and extensive to demand a separate work; these subjects were originally connected with the Theory of Probability but may now be considered to form an independent kingdom in mathematical science . . .

Nor is Todhunter alone in his misgivings. Though intensive study of algebraic probability has been compulsory till recently as a vocational discipline only in the training of Actuaries, and still occupies a place of honour in the actuarial curriculum, an influential body of actuarial mathematicians subscribe to the conclusion last stated, as the reader may infer from the pages of the *Journal of the Institute of Actuaries* (1945). A report on papers written for the Twelfth International Congress of Actuaries with the heading *The Theory of Probability in Insurance*, expresses the view:

The part that the theory of probability plays in insurance is regarded as important by Baptiste, de Finetti, Berger, doubtful by Hammon and Clarke, Lah, and of little consequence by Hagstroem, Shannon, ten Pas.

The most we can say about the relevance of the classical theory to the task of the Actuary is that the arithmetical operations which we may carry out with confidence as a prescription of long-term policy when we can legitimately presume that Uspensky's two-fold condition does hold good suggest a rough and ready pattern for short-term policy in a changing situation when the firm fixes its stake with a sufficiently generous safety margin. The fate of the *Equitable* indeed vindicates the merit of such a margin.

The *Equitable* felicitously based its early premiums on the Northampton Life Table, the author of which grossly underestimated the population at risk by using the data of baptismal registers in a township with a large dissenting community. By this error the firm added a vast sum to its capital assets. By the same error the British government at that time lost heavily on annuity outpayments to its pensioners. If the actuaries of

the *Equitable* did in fact shape its policy in conformity with Bernoulli's theorem and with Bernoulli's own recipe for a *posteriori* ascertainment of the putatively constant probability of the event at each trial, the efficacy of their recommendations confers its sanction neither on the theorem nor on the recipe.

How powerfully the analogy between insurance risks and those of the gambler was to influence statistical thinking, the now technical connotation of the word *expectation* sufficiently attests. A remarkable contribution of D. Bernoulli (1760) contemporaneous with the *Recherches* of Euler and entitled *Essai d'une nouvelle analyse de la mortalité causée par la petite Vérole*, etc., provides a direct link between actuarial preoccupations and Quetelet's identification of the so-called laws of chance with the laws of population. At a time when the Turkish practice of inoculation had newly reached Western Europe, the vital statistics of smallpox had already enlisted the attention of D'Alembert. In the *Essai* mentioned, D. Bernoulli set out how the incidence of a disease (*smallpox*) which confers immunity, will decline as age advances. Though this, for its time so remarkable feat of analysis, seems to have passed unnoticed by British epidemiologists till Greenwood drew our attention to it, it may well have provoked interest on the Continent, where there appeared in 1840 the first noteworthy book about medical statistics, namely *Principes généraux de Statistique Médicale* by Jules Gavarret. Bernoulli's essay on smallpox is itself justly noteworthy as a landmark in the history of epidemiology; but it can stand on its own feet without the crutches of a stochastic calculus to support its weight.

The formal identification of the risk of dying with the risks of casting a double six in two tosses of a die becomes more explicit when Laplace introduces the entirely Platonic notion of the infinite hypothetical population, expounded by Todhunter as follows:

Laplace then considers the probability of the results founded on tables of mortality: he supposes that if we had observations of the extent of life of an infinite number of infants the tables would be perfect, and he estimates the probability that the tables formed from a finite number of infants will deviate to an assigned extent from the theoretically perfect tables.

Two new concepts emerge in the treatment of the problem by Laplace. Initially and explicitly, he asks us to conceive an infinite and hypothetical population of which any actual population is an imperfect sample. Next, in effect, he asks us to conceive our observed sample as a sample taken randomwise therefrom. The assumption last stated is the kingpin of the theory of statistical inference expounded by R. A. Fisher, who follows in the tradition of Karl Pearson when he asserts (1925) unequivocally:

any body of numerical observation or qualitative data, thrown into a numerical form as frequencies may be interpreted as a random sample of some infinite hypothetical population of possible values. (*Proc. Camb. Phil. Soc.*, XXII.)

To do justice to the *infinite hypothetical population* it will be useful to examine separately the implications of its infinite contents and its hypothetical attributes. Its model was the urn of the Royal Lottery. It is not wholly an accident of fate that the urn model invoked in the promulgation of the doctrine of Laplace was also the funeral urn of the Forward Look. No familiar conventions of daily life inhibit an inclination prompted by a theoretical prepossession to enlarge both its size and the number of *billets* or balls it contains; but a so seemingly innocent materialisation seemingly also confers the licence to dispense with the temporal series of the event, as we conceive it for the benefit of the gambler who perseveres in quest of his winnings. Our universe of choice is a *bran tub universe* into which each of an infinitude of children is at liberty to dip and to withdraw therefrom a single  $r$ -fold sample at one and the same instant of time. Thus the Forward Look has now become a fixed stare. The framework of repetition is no longer historic; and its conceptualisation in a static imagery invites us to forget an essential property of the new model. If our model justifiably endows the event with constant probability at each trial, what we predicate of such sampling will be more or less relevant to what will happen at next year's Christmas Party only if the bran tub of today is in all relevant particulars identical with the bran tub of tomorrow.

Before we can persuade ourselves that any sample from this

bran tub universe is indeed what Fisher calls a random sample, and hence that we can speak of the constant probability of the event at each trial, we encounter a difficulty discussed in an earlier chapter (p. 75). The mere fact that our card pack is infinite does not dispense with the need to prescribe a rule to ensure randomwise removal of samples therefrom. If we concede that the observed population of the village of Glynceiriog in the year 1953 is in some transcendental sense a sample of some infinite hypothetical population, the assumption that the latter is also, as such, an unchanging universe of choice does not suffice to endorse the relevance of the classical calculus to the process of sampling therefrom. We must equip it with all the essential properties of the *Irregular Kollektiv*; and it is by no means easy to discern a single adequate reason for doing so. Kendall, who subscribes to the enlistment of a new term with a so highly emotive content when asserting (p. 19) that "the population conceived of as parental to an observed distribution is fundamental to statistical inference," likewise concedes with commendable candour that the concept of the infinite hypothetical population:

. . . is not required (and indeed has been explicitly rejected by Jeffreys) in the approach which takes probability as an undefinable measurement of attitudes of doubt. But if we take probability as a relative frequency, then to speak of the probability of a sample such as that given by throwing a die or growing wheat on a plot of soil, we must consider the sample against the background of a population. There are obvious logical difficulties in regarding such a sample as a selection—it is a selection without a choice—and still greater difficulties about supposing the selection to be random; for to do so we must try to imagine that all the other members of the population, themselves imaginary, had an equal probability of assuming the mantle of reality, and that in some way the actual event was chosen to do so. This is, to me at all events, a most baffling conception. (*The Advanced Theory of Statistics*, Vol. I.)

The popularisation of the term population for what we may now call the *Irregular Kollektiv*, with less risk of the self-deception incident to use of common speech in a highly technical context, was largely due to Quetelet, whose *niche* in the history of statistical theory we have already noticed (pp. 18-20). Quetelet

wrote in the context of Gauss, and we must therefore defer a full consideration of the misconceptions embedded in the *mystique* he bequeathed to posterity. As they concern the theme of this chapter, what is most relevant emerges from the following citation from Keynes who elsewhere remarks "there is scarcely any permanent accurate contribution to knowledge which can be associated with his name":

Quetelet very much increased the number of instances of the Law of Great Numbers and also brought into prominence a slightly variant type of it, of which a characteristic example is the law of heights, according to which the heights of any considerable sample of a population tend to group themselves according to a certain well-known curve. His instances were chiefly drawn from social statistics and many of them were of a kind well calculated to strike the imagination—the regularity of the number of suicides, *l'effrayante exactitude avec laquelle les crimes se reproduisent* and so forth. Quetelet writes with an almost religious awe of these mysterious laws, and certainly makes the mistake of treating them as being adequate and complete in themselves like the laws of physics and as little needing any further analysis or explanation. Quetelet's sensational language may have given a considerable impetus to the collection of social statistics but it also involved statistics in a slight element of suspicion. . . . The suspicion of quackery has not yet disappeared. (*Treatise on Probability*, p. 335.)

In this context, we must interpret the *Law of Great Numbers* in an empirical sense. Experience commonly shows that social indices of a large population fluctuate less widely than those of small ones as will commonly be true if: (a) the latter are its constituents; (b) they do not all fluctuate in the same direction at the same time. We can imagine many reasons for this, if we fix our attention on any particular rate and explore the secular or local operation of known agencies which determine its magnitude; but there is no *prima facie* reason for regarding a principle of statistical equilibrium formulated in such terms as an outcome of circumstances to which the so-called laws of large numbers severally identified with later elaborations of Bernoulli's theorem by Poisson, Tchebyshev, Cantelli and Markhov are pertinent. Nor did Quetelet himself rely on the

algebra of Bernoulli or of Poisson to interpret it in stochastic terms.

We can do full justice to Quetelet's reasoning, only if we first acquaint ourselves with the postulates and proper scope of the Gaussian calculus. The well-known curve referred to by Keynes is indeed the Gaussian Curve of Error, thereafter endorsed as the *normal* distribution; and Quetelet's major error arises from a still too widespread belief that we can legitimately infer an interpretative law of nature from the applicability of an algebraic formula to the contour of a fitting curve. Fortunately, this misconception was less prevalent at the beginning of the nineteenth century, when the formal equivalence of the partial differential equations for conduction of heat and for diffusion of fluids might otherwise have endowed the *caloric* with a longer lease of life. At an early stage, the student of physics now gratefully welcomes the same familiar graphical representations of metrical relations in a diversity of otherwise dissimilar physical phenomena with no disposition to draw false or gratuitous conclusions about what they may have in common; but Quetelet's eagerness to descry curves which recall sampling distributions assumed as the basis of the Gaussian theory of errors in a range of studies less attuned to such coincidences gave a powerful and lasting impetus to the gratuitous intrusion of a stochastic theory of errors of observation into the domain of structural variation. The confusion is still with us. Thus R. A. Fisher, in a passage elsewhere cited (p. 501) from the *Design of Experiments* (p. 195), refers to the *theory of errors* in a context which unequivocally identifies the relevant distribution of a test procedure with that of an infinite population on all fours with that of Quetelet's so-called law of heights. Even more explicitly (p. 56 *op. cit.*) he declares that

the mean square corresponding to 28 degrees of freedom ascribed to *error*, is available as an estimate of the variance of a single plot due to the uncontrolled causes which constitute the errors of our determinations.

If we take this step, we have in fact transferred the responsibility for consistent adherence to the rules of the classical calculus of probabilities from the shoulders of the Chevalier de

Méré to those of Nature disguised as the Divine Dealer. Against the background of his usually urbane comments on other writers, his subsequent criticism of the Galton-Pearson measure of correlation in social enquiries and his final conclusions that the calculus of aggregates alone remains intact amidst the wreckage, the ferocity of Keynes when dealing with Quetelet thus becomes intelligible and fully consistent with his antagonism to the then nascent claims of statistical inference in the domain of the social sciences. For the classical framework of repeated trials is not the way in which Nature makes history.

In the domain of the deductive genetical theory of population Quetelet's much-quoted metaphor about Nature's urn need not lead us far astray; and we can invoke a stochastic interpretation of erroneous observation in the experimental sciences without excessive violence to the classical tradition. In laboratory work, we can postulate a protracted framework of repetition on the explicit assumption that we have the relevant variables under control; but in large measure, the subject-matter of the social sciences consists of unique historical situations which (as such) are unrepeatable in any sense unless we regard the shadow world of human experience as a sample of an infinite stock laid up in the Platonic heaven of universals. Anything we do in the laboratory of government changes the structure of the apparatus and the materials involved in the experiment. The bran-tub universe is an illusion, because the bran-tub at the Christmas party next year will be a new one.

None the less, we cannot concur with Keynes in dismissing Quetelet as an addle-pated and rhetorical enthusiast without likewise condemning a generation which embraced his *mystique*. To understand its appeal, we must recreate the intellectual climate of the period. We have then to remind ourselves that the discussion of annuities in the classical context had attuned the ears of the generation to whom Quetelet and de Morgan addressed themselves to the acceptance of a stable rate of birth and death as a law of Nature proclaimed with the fervour of the pulpit by the parson Malthus and endorsed by Darwin with the status of an article of faith. How else can we explain that all extant books in actuarial practice exhibit the simple arithmetic of the construction of a life table, which is in fact a summary of



current events in a rapidly changing situation, as an exercise in probability in contradistinction to an exercise in the Rule of Three? Those of us who are near sixty take for granted a steadily falling infant death rate over the period of a lifetime; and reliable statistics inform us that the mean duration of life in our community has increased steadily since 1837. None the less, the craft guild of actuaries still maintains its privileged position by imposing on candidates for admission a discipline which has had little relevance to the *modus operandi* of the insurance business since the office of the Government Actuary came into being.

It is therefore a circumstance of no mean significance that Karl Pearson's school, which built on the foundations laid by Quetelet and by his disciple Galton opposed with every spurious syllogism endorsed by a basically false theory of inheritance the claim that considerable improvement in the health of the community unaccompanied by selective breeding could be other than an ephemeral event. On any other view, it is impossible to conceive *successive* statistics of a population as samples extracted from one and the same *Irregular Kollektiv*. The tenacity of Pearson's belief in the stability of human populations is thus easy to understand, if difficult to condone. It re-asserts itself today as a deeply religious anxiety to resurrect the doctrine of the gloomy parson after its ceremonial crucifixion in the writings of R. R. Kuczynski and Enid Charles in so far as it has any relevance to population growth in affluent societies favourable to the spread of contraceptive practice.

Kendall concludes the passage last cited from his works by remarking:

At the same time, it has to be admitted that certain events such as dice-throwing do happen as if the constituents were chosen at random from an existent population, and it accordingly seems that the concept of the hypothetical population can be justified empirically.

We may indeed concede that the concept is meaningful and tailored to the requirements of a stochastic calculus in the domain of American dice which behave like the dice of Uspensky (p. 53). If so, we then conceive the infinite popula-

tion in terms of an historic framework of endless repetition; and the stochastic propriety of the behaviour of the die will suffice to justify the concept as "fundamental" in the wider domain of statistical inference only in so far as we can justifiably endow natural phenomena with its relevant properties. In particular, and needless to say, we implicitly concede that the probability of the event remains constant at each trial. How Quetelet accomplished the feat of proving to the satisfaction of his contemporaries that variation in the domain of biological and sociological enquiry is indeed the manifestation of a self-randomising process, we shall examine more closely in the *Gaussian milieu*. Here it suffices to remark that the exclusion of this principle, as a precondition to the tie-up of the algebraic law of great numbers with the empirical stability of large units of population, also and inexorably excludes from the proper domain of stochastic theory a theme which has lately become another fashionable playground for algebraic exploits. A single quotation will suffice to show that the present writer is not alone in affirming this conviction. In *The Theory and Measurement of Demand* (pp. 214-15), Henry Schultz (1938) declares:

Now time series, especially those relating to social and economic phenomena, are likely to violate in a marked degree the fundamental assumption . . . that not only the successive items in the series but also the successive parts into which the series may be divided must be random selections from the *same* universe. Time series are, in fact, a group of successive items with a characteristic conformation. Such series . . . cannot be considered as a random sample of any definable universe except in a very unreal sense. Nor are the successive items in the series independent of one another. . . . The fact is that the "universe" of our time series does not "stay put," and the "relevant conditions" under which the sampling must be carried out cannot be recreated. . . . It is clear, then, that standard errors derived from time series relating to social and economic phenomena do not have the same heuristic properties that they have, or are supposed to have, in the natural sciences.

Aside from a legitimate objection to the licence conferred on gratuitous extension of stochastic principles to unique historical situations by the use of the word population ambiguously both for what we have elsewhere called the fixed universe of choice

and for the dynamic orderless collective of von Mises, its prevalence in statistical literature is exceptionable because it has added a new difficulty to the exposition of the theory of probability by depriving a *frequency distribution* of any single clear-cut meaning. Thus current statistical textbooks exhibit results of experiments on tossing dice side by side with records of rainfall and weights of beans. In one and the same context, a frequency distribution may thus mean:

(i) a precise specification (elsewhere called a unit sample distribution) of actual or relative frequencies of discrete score values in either an infinite or a finite known universe of choice such as an urn;

(ii) a comparable specification of relative frequencies in a hypothetical continuum of score values;

(iii) a deductive algebraic specification of the relative frequencies of an infinitude of  $r$ -fold samples taken randomwise from the appropriate universe of choice, if specifiable in terms of (i) or (ii);

(iv) an empirical specification of numbers or proportions of individual numbers of any finite assemblage classifiable with respect to some numerically specifiable attribute by a system of scores which may stand for counts (e.g. deaths per thousand, red blood cells per *cmm*) or for measurements (e.g. heights) grouped in discrete intervals;

(v) a descriptive curve of best fit for (iv) supposedly exhibiting the composition of a parent assemblage w.r.t. which (iv) itself is a sample.

It is therefore all too easy for the beginner to assume that a hypothetical construct such as (ii) can rightly claim the same factual credentials as (i) or that (iv) stands in the same relation to (v) as (iii) to (i). That the legitimacy of such assumptions is highly debatable will have sufficiently emerged from our examination of the classical heritage. Whether considerations which emerge in a later context can substantiate them will dictate a rational assessment of the enduring contribution statistical theory can make to the advancement of science.

When we do examine in greater detail the arguments Quetelet advanced to sustain the thesis that empirical distributions such as the so-called law of heights are the outward and

visible sign of a natural shuffling process, the circumstance that he was the foremost Belgian astronomer and meteorologist in the thirties of the nineteenth century will disclose more than one intelligible clue to the genesis of his monumental *non sequitur*. Here it will suffice to refer to the attitude Quetelet adopted to scientific law. By no means the amateur and the playboy portrayed by Keynes, Quetelet was indeed an academician of high esteem. As founder of the Belgian national observatory, he was a custodian of the Newton-Laplace cosmogony with a peculiar interest in natural periodicities, meteorological and ecological. If we subscribe both to the view that some unique verbal formulation can embrace every situation in which men of science speak of a law of nature and to the eighteenth-century belief that Newton's law of universal gravitation is its supreme paradigm, what experimental science must now dismiss as an absurdity assumes the aspect of a truism in the highly respectable tradition extending from Aristarchus and Hipparchus to Ptolemy, from Ptolemy to Kepler and from Kepler to Einstein.

In the proper domain of celestial periodicities beyond man's power to control, the identification of the discovery of such a law with an exercise in curve-fitting is consistent with what we do in fact commonly agree to call a law of nature in the same context. In asserting Kepler's laws of the planets as in asserting Calvin's laws of God, we ourselves occupy the role of passive spectators impotent to set aside what the Almighty predestined and foreordained in the beginning. That all scientific laws are expressible in such terms is likewise consistent with a social doctrine which extends from Kepler's contemporary Calvin through Adam Smith and Malthus to Darwin, to Galton, the father of the political cult variously named *eugenics* or *Rassenhygiene*, to Galton's disciple Karl Pearson and to Dr. Malan. Thus Quetelet, Malthus and Galton alike conceived a law of society in terms entirely consistent with the prevailing concept of natural law as the statement of an unchangeable regularity which man's own frail and sinful nature can never gainsay.

Though himself confessedly an impenitent devotee of Malthus, Keynes derides the awe with which Quetelet proclaims these eternal verities; but a concept of natural law so widely acknowledged, and one which Karl Pearson took as the

text of his *Grammar*, has inspired others of the same persuasion, notably Galton, to utterances equally in tune with those of Calvinistic theologians proclaiming the inscrutable exclusion of all the sons of Ham from the benefits of the New Dispensation of Grace. If one endorses the identification of all scientific law with statements on all fours with charts of the unchanging periodicities of the heavenly bodies, one can complain with little justice about the melancholy earnestness of Quetelet's declaration:

We pass from one year to another with the sad perspective of seeing the same crimes reproduced in the same order and calling down the same punishments in the same proportions. Sad condition of humanity. . . . We might enumerate in advance how many individuals will stain their hands in the blood of their fellows, how many will be forgers, how many will be poisoners, almost we can enumerate in advance how many births and deaths there should occur. There is a budget we pay with a frightful regularity; it is that of prisons, chains and the scaffold.

Quetelet had at least one excuse which we cannot plead in exoneration of Galton's equally rhetorical relapses. All the thinking in his *Essai de Physique Sociale* had already taken shape before a laboratory demonstration of the first law of thermodynamics by Joule vindicated the common sense of British engineers in the following of James Watt. Here, admittedly, is law conceived in terms no less inexorable than the law of universal gravitation; but the intention is wholly different. The emphasis is henceforth on law conceived as a recipe for action in man's exercise of his power to make all things new. Quetelet's teaching also antedated the controversy over evolution. Here we must adjust our sights again. Like the law delivered on tables of stone, what is law is now the written word of the record of the rocks. Eventually, we have thus three concepts of law to accommodate, that of the Nautical Almanac, that of the Cookery Book and that of the Statute Book. Whether J. S. Mill or Karl Pearson succeeded in finding a formula for scientific method able to subsume all three or even any two of them is still debatable.

Perhaps the worst we may say of Quetelet is that he was consistent in the worst sense of the term. No less than his

professional preoccupation with a prescription of natural law dictated by the teaching of the Newtonian era, Quetelet's nostalgia for the *ancien régime* in the stormy setting of 1848 is the hallmark of one of two conflicting ideologies which reappear in a new guise at times congenial to innovation. We fail to recognise the peculiar strength of its appeal to the individual, if we disregard the vigour with which many innovators of the past have assumed the role of champions of ancient liberties. In different places and at different times the partisans of over-privilege and the pamphleteers of the have-nots may vigorously espouse a dogmatic determinism or a euphoric libertarianism for reasons to which logic alone furnishes no clue.

In Quetelet's boyhood a united front against clerical authority had little inclination to recognise a dichotomy, which will emerge again and again as we proceed with our examination of current statistical doctrine. It does not assert itself in the contemporary writings of Whewell and Mill. It could enlist little attention in the climate of the third session of the British Association, when Quetelet brought the evangel of *Physique Sociale* to Cambridge. In the following of the *École Polytechnique*, the exhilaration of successfully imposing on every admissible branch of human knowledge a discipline congenial to the cosmogony of Laplace and Lagrange was a sentiment shared by strange bedfellows. Of necessity the undertaking had to accommodate itself to two streams of humanistic thinking. In the setting of the 18th *Brumaire*, Comte is the pilot of one, Quetelet of the other. The successors of both are with us; but few among those whose preferences lean to the conservative and descriptive rather than to the perfectionist and experimentalist tradition would now care to subscribe to the rationale of Quetelet's conservatism. We still know little about how to diminish crime; but a programme of research with that end in view is by no means meet for contemptuous dismissal as an unscientific aspiration engendered in the enthusiasm of the English Evangelical revival and the French Revolution.

We do not err too widely from the track of our assigned enquiry, if we thus pause to see Quetelet and his enthusiastic following in the milieu of the 1848 Commune and the Great Exhibition of 1851. As the parent of what we have elsewhere

called the *calculus of exploration*, he was more than the inventor of a technique. He was the architect of a system of values and of an epistemology later inflicted by Karl Pearson on a generation still surviving. The intellectual climate of Quetelet's generation thus forces on our attention an issue which we shall have to face in more than one form, if we intend to carry out an exhaustive revaluation of the claims and credentials of current statistical theory. At a later stage, we shall see how prominent a part Pearson's unique formula for enquiry subsumed by the expression *scientific method* and by any unique definition of *law* suitable to all the uses of scientific enquiry plays in the background of the theory of regression. If we hope to pass judgment with wisdom on conflicting current claims of statistical inference, we shall also find ourselves forced to re-examine Mill's attempt to disclose a common denominator for the reflective and retrospective disciplines on the one hand and for the activist and prospective disciplines on the other.

## CHAPTER FIVE

### THE BACKWARD LOOK AND THE BALANCE SHEET OF THOMAS BAYES

WE MAY DIVIDE what we have called the classical period into two phases. The first culminates in the publication of the *Ars Conjectandi*. At this stage, the calculus of probability makes only one claim to topical relevance. Inasmuch as it endorses a rationale for the division of stakes in games of chance, it can supply to court circles a new theme for conversational entertainment; but it can offer to a thrifty bourgeoisie no certain recipe for swelling a bank balance. It has as yet no solid foothold in the world of affairs. Were it not for what followed shortly after, it would have accomplished little to engage the interest of posterity.

In the background of the second phase we discern the juncture of two circumstances. An effete dynasty, otherwise unable to endow the whims of its mistresses from a treasury which supported the largest standing army of the time, had successfully commercialised the art of gambling to beguile the surrender of their savings from subjects sullenly rebellious against the salt tax. Meanwhile, the practice of life insurance had gained a firm foothold with new prospects of gain for merchants of substance. In this setting, there was a new public eager enough to give ear to the popularisation of the theory of probability. Diderot's co-editor of the greatest popular work of all time was none other than d'Alembert, himself a prominent exponent of the theory.

If the identification of insurance risks with the hazards of the gaming table is exceptionable from our own viewpoint, it committed the contemporaries of d'Alembert to no rupture with the doctrine of their predecessors. The latter had relied exclusively on their own intuitions to prescribe when the concept of *random* choice is admissible. If those who succeeded them indulged in the same liberty, they did not violate the explicit teaching of their teachers. To be sure, a generation less tolerant of Platonic notions might have rejected the infinite



hypothetical population of Laplace as a gratuitous metaphysical abstraction; but its intrinsically stochastic properties will be repugnant to our inclinations only if we have accepted the obligation to define circumstances relevant to our recognition of random occurrence. We have already seen that no writer on probability before von Mises paid much attention to this issue; and what now seems to us to be a formidable factual obstacle to the acceptance of the infinite hypothetical population in the context of life insurance provoked no remonstrance from the many adherents to Quetelet's belief in a fixed norm about which births and deaths fluctuate in accordance with a law comparable to the Gaussian law of error.

In short, de Moivre, Euler and d'Alembert and D. Bernoulli did not overtly deviate from the *Forward Look* of their predecessors, when they turned their attention to the theory of the Life Table. What signalises a manifest break with the past is the announcement of the doctrine of *inverse probability*. The principal proponent of the doctrine, commonly associated with the name of an English dissenting divine in the circle of Joseph Priestley, was Laplace himself; but no account of it is complete if we fail to mention a theorem, which has kept posterity guessing for nearly two centuries. In 1763 the Royal Society published in its *Philosophical Transactions* a posthumous contribution of the Rev. Thomas Bayes. Dr. Richard Price, himself like Bayes and Priestley a Unitarian minister and like the latter a Fellow, communicated it to the Society. Price prepared for publication the script from the author's unrevised relict. Were it not for the fact that Laplace later acknowledged it as the spiritual parent of his own *mystique*, few of us would have heard of it. As matters stand, most subsequent writers on statistical theory up to and including the present time recognise the *Essay towards solving a Problem in the Doctrine of Chance* as a landmark, and its contents as a challenge or programme, according to taste.

Price merits comment *en passant* on his own account. From a *View of the Rise and Progress of the Equitable Society* (1828) by its actuary, William Morgan, F.R.S., we learn that the promoters of the *Equitable*, when first incorporated (1762), gained much "profit by the advice and instruction of such a person as

Dr. Price" who "communicated to the court of directors some observations on the proper method of keeping the accounts and determining from year to year the state of the society." We also read that

this invaluable communication contained three plans for that purpose detailed at considerable length:—the first, by ascertaining the proportion of the claims to the premiums; the second, by comparing the decrements of life in the Society with those in the Table, from which its premiums were computed; the third, by making a separate computation of the values of all the different policies of assurance, and comparing the amount with the capital of the Society. In addition to these plans, Dr. Price, among other important advice, urged the necessity of altering the tables of premiums then published in the *Short Account* of the Society, not only as being exorbitant, but absurd, and inconsistent with the result of all observations—alluding particularly to the female and youth hazards. These extraordinary charges were, in consequence, immediately abolished, and each of the three plans above mentioned was adopted for ascertaining the state of the Society from the year 1768 to the year 1776 inclusive. By the first of these plans it appeared that, on an average, during the nine preceding years, the annual surplus had been about 3000 £. By the second plan, that the probabilities of life in the Society had been higher than those in Mr. Dodson's Table, from which its premiums were computed, in the proportion of three to two. And by the third plan, that the whole surplus stock amounted nearly to 30,000 £. In consequence of results so highly favourable to the Society, the premiums were reduced *one-tenth*: which does not, however, appear to have had any great effect, either in increasing the business or in lessening the annual surplus; for the continual accession of new members, by adding to the number of the old ones, fully supplied the deficiency produced in the surplus by the reduction of the premiums, and thus made it increase very nearly in the same proportion as in the two or three preceding years.

It appears that Price was directly responsible for the adoption of the Northampton Life Table already mentioned (p. 96).

In the year 1780, Dr. Price had formed a great number of tables deduced from the probabilities of life at Sweden, Chester, Northampton, and other places, preparatory to the fourth edition of his work

on Reversionary Payments. These tables he considered as more correct than any hitherto published, and recommended the adoption either of the Chester or the Northampton to the Society, in lieu of the very imperfect table from which its premiums had hitherto been computed. This, like every other measure recommended by Dr. Price, was agreed to without hesitation, and before the end of the year 1781, a complete set of tables was formed from the Northampton observations, consisting of more than 20,000 computations, and containing the values of single and joint lives of all ages, and the single and annual premiums of assurance of every description; but the latter, though computed at 3 per cent, were so far below the premiums then in use, that it was thought proper to make an addition of 15 per cent to them, to prevent too sudden a reduction in the annual income of the Society. By the adoption of these new tables, the annual premiums, which would then have been 36,000 £, if the old tables had been continued, were reduced to little more than 32,000 £, and in order to compensate the members then existing, for having contributed to the success of the Society by the payment of higher premiums than were necessary, an addition was made to each 100 £ assured by them of thirty shillings for every payment which had been made prior to the 1st of January 1782. During this and the three following years, the number of new assurances annually increased about *one half* of their former number, and the annual income in the same proportion. This rapid growth of the Society, added to the circumstance of no particular investigation having been made of its actual state since the year 1776, led to the resolution for making a fresh investigation before any measures should be adopted that had a tendency to affect the finances of the Society. In the course of the year 1785, this laborious work was accomplished; and the result proved so highly favourable that it was determined not only to take off the charge of 15 per cent on the premiums, but to make a further addition of 1 to each 100 £ for every payment made prior to the 1st of January 1786. By these operations the surplus of 164,000 £ was reduced to 110,000, and every person assured prior to 1772, had 30 per cent added to the sum originally assured.

To any careful reader of Morgan's pamphlet, it will be clear that Dr. Price was not the obsessional gambler who adheres to a rule regardless of the consequences; and the hard-headed business men who benefited from his advice and instruction had no scruples about changing the rules of the game as good

luck enjoined. A break with the classical tradition was therefore inevitable, if the stochastic calculus was henceforth to annex a territory with so promising a prospect of full employment for the algebraist, and the more so when the immediate successors of Price began to entertain misgivings about the constancy of the *billets* in the mortality lottery. It was soon clear that a malignant fate changes the contents of the funeral urn :

Ever since the promulgation of Mr. Malthus's system, a general alarm has been excited among all ranks and conditions of men in the United Kingdom, that the population increases so fast, and the life of man is extended to such a length, that the fruits of the earth will soon be insufficient to preserve us from perishing by famine. . . . The tables also which formerly denoted the probabilities of life, are now said to be no longer applicable to the improved health and constitutions of the present race, a circumstance which is rendered the more remarkable from the acknowledged increase of pauperism among the greater number of them.

In the first Report of the Committee of the House of Commons on Friendly Societies, a large collection of tables is inserted ; which, if we judged from the long line of decimals to which the values are extended, might be considered as a work of uncommon accuracy, and founded on documents which did not even admit of error. But of these documents we have not sufficient information, nor is it indeed a matter of much consequence—the tables themselves, though computed to the millionth part of a *farthing*, being so wrong in the *pounds*, especially in the case of female lives, as to deserve little or no regard. By the assistance of these tables, the notable discovery has been made of the loss sustained by the public of many thousands every week, for several years past, by granting annuities on lives, computed from the Northampton Table of Observations, which had the surprising effect of alarming the House of Commons into a vote for immediately repealing the law which authorized that measure.

The document is worthy of study in its entirety, if the reader still entertains any illusions about the possibility of regulating the affairs of a successful life insurance corporation by strict adherence to the principles of J. Bernoulli. *Inter alia* it is a useful source of information about the early practice of life assurance. Before 1762, the *Amicable* “which had existed from

the beginning of the century was the only society formed for the express purpose of making assurances on single lives. . . . Although the *Royal Exchange* and the *London* assurance office were empowered by their characters to assure lives, they seldom availed themselves of that power." It does not appear that Dr. Price enriched himself in his capacity as consultant to the Equitable. Nor is it likely that he did so. In his time, the Anglican Church was still the custodian of what social security the masses enjoyed; and dissent could successfully challenge the prerogative of the parish council as the dispenser of public charity only by showing that thrift pays in the long run. Henceforth paid-up premiums conferred on the *paterfamilias* of the non-conformist household a certificate of piety. As the familiar name of the *Wesleyan and General* reminds us, the chapel vestry remained the recruiting station of the insurance company throughout the nineteenth century. We may therefore charitably assume that the advice and instruction by which the court of the Equitable derived such signal profit, no less than the work of editing the mathematical relict of his ministerial colleague the Rev. Thomas Bayes, was a labour of love.

Though Bayes's theorem and Bayes's postulate or scholium meet the eye in every contemporary controversial contribution to theoretical statistics, it will rarely happen that three statisticians chosen randomwise will wholly agree about what precisely Bayes did say. Nor will they necessarily agree about which of two different propositions constitute his theorem. In justice to Bayes, one may say that one form of the theorem attributed to him certainly does not occur in his works, and the other, which is a modern interpretation of the *ipsissima verba*, is dubiously consistent with the author's intentions. Such misunderstanding would be merely of lexicographical interest, were it not also true that later generations have chosen to identify, fairly or falsely, the views of Laplace with those of Bayes himself; but we need not wonder why there can be so much confusion about the issue when all the relevant source material is accessible in any creditable university library. Bayes employed in his own notes a symbolism of an older vintage than that of his French contemporaries, being himself steeped in the Newtonian method. That his language is excessively obscure, we may

tolerantly condone, since he had no opportunity to assemble the material for publication. Nor need we blame his literary executor who confessedly indulged himself so freely in footnotes and addenda not necessarily consistent with the author's intentions.

However, these circumstances do not throw much light on one novel feature of the memoir, and one which has also been contributory to subsequent mystification. Pascal, pre-eminent among the founding fathers, was likewise the author of a treatise on discrete figurate number series which have a special significance in the elementary calculus of choice; and it would be fair to say that the notion of a continuum, except as a background for performing computations which would otherwise be intolerably laborious, does not intrude aggressively in the classical period. In the *Essay* of Bayes, we are no longer counting discrete pips and discrete faces. We find ourselves in a non-enumerable domain of Platonic points, where summation is quadrature in the most literal sense of the term. In defiance of the, not as yet established, first law of thermodynamics, perfectly spherical and perfectly smooth billiard balls roll at the behest of the author's pen on frictionless planes as the theme of Euclidean demonstrations in the grand manner of the *Principia*.

It is possible to terminate otherwise unending theological disputation about what is real Christianity or about what Karl Marx really meant, if one restricts the opportunity for self-indulgence in phantasy by also limiting the field of discourse to what the gospels actually record or to what *Das Kapital* does indeed state. We shall therefore do well to study closely the text of the *Philosophical Transactions* for the year 1763, interpreting the *ipsissima verba* with due regard to the idiom of the author's contemporaries. The posthumous memoir has two parts. The first sets out as Euclidean propositions, in a style still current when Loney's textbooks of dynamics, statics and hydrostatics became obsolete long after my own schooldays, a few school certificate level tautologies of the classical theory. It is notable only because one of them, one which prompted the editor to comments seemingly inconsistent with the author's intention and one which prompts the first historian of probability to bewildered reflection on its ostensible novelty and

possible self-evidence, plays an important role in the section which follows. It there provides a platform for a highly sophisticated pun; and its examination will help us to see the pitfalls of a symbolism which does service to those who locate probability *in* the mind.

To convey the gist of Proposition V, we may suppose that we have several dice to toss so many ( $r$ ) times or several urns from which to extract randomwise so many ( $r$ ) balls, recording the score of an  $r$ -fold trial as  $x$ . One of these dice or one of the urns we shall call an urn of type H. If we follow an appropriate recipe to ensure randomwise choice of urn or die, we may state comprehensively the elementary theorem of multiplication of probabilities for a situation involving *two* acts of choice by recourse to appropriate symbols as follows:

- $P_{h,r}$  is the unconditional probability that we *shall* first choose an urn or die of type H;
- $P_{x,r}$  is the unconditional probability that the score of the  $r$ -fold trial *will* be  $x$ ;
- $L_{x,hr}$  is the conditional probability that the  $r$ -fold trial score *will* be  $x$  if we do in fact first choose an urn or die of type H;
- $L_{h,xr}$  is the conditional probability that we *shall* choose an urn or die of type H in the infinite sub-set of  $r$ -fold trials which yield the particular score  $x$ ;
- $P_{h,xr}$  is the unconditional probability of the compound event that we shall both score  $x$  and choose an urn or die of type H.

The elementary multiplication theorem then takes the comprehensive form

$$P_{x,r} \cdot L_{h,xr} = P_{h,xr} = P_{h,r} \cdot L_{x,hr}$$

Thus we may write with equal propriety:

$$L_{h,xr} = \frac{P_{h,r} \cdot L_{x,hr}}{P_{x,r}} \text{ or } L_{x,hr} = \frac{P_{x,r} \cdot L_{h,xr}}{P_{h,r}} \quad (i)$$

Now there is nothing new in the identity on the left, which is in fact Proposition V of Section I of the Bayes memoir; but one can read into it something foreign to the thought of the classical

period in the light of what use Bayes makes of it in Section II. There the author advances, intentionally or otherwise, a new idea. We think of multiplication as an operation to which division is the corresponding *inverse* operation; and our algebraic jargon here becomes entangled in our daily habits of verbal discourse. Bayes may have thought, and Price certainly did think, that the division theorem set forth as (i) above confers the licence to look backwards as well as forwards, when we specify a rule of procedure within the framework of the classical theory of risks in games of chance; but the intentional insertion of the italicised future auxiliary in the foregoing specification of our symbols should suffice to remind us that this is a verbal trick. From his own words, which Todhunter complains of as obscure, there is no indication that Bayes intended to exploit the punning potentialities of a trivial tautology for rhetorical effect in this context. It may help to clarify what follows if we here recall his own idiom:

If there be two subsequent events, the probability of the 2nd  $b/N$  and the probability of both together  $P/N$ , and it being first discovered that the 2nd event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is  $P/b$ .

The use of the verb *guess* by Bayes in this context is instructive. It recalls the literal derivation of a word current in English long before it became part of the jargon of statistical theory, as when Dean Swift\* writes: "I am master of the stockastic art; and by virtue of that, I divine that those Greek words in that discourse have crept from the margin into the text otherwise than the author intended." In the classical theory of wagers, every bet is a guess; and the probability that the guess will be right is the probability that the event specified by the guess will occur in the endless sequence of the event. In the classical tradition, we can assign a probability to our guesswork in this sense only if we confine our statements to guesses dictated by a rule stated in advance; and the use of the past tense in this citation entails no explicit renunciation of the

\* In *Right of Precedence between Physicians and Civilians enquired into*. Misc. Works. 1720.



Forward Look, if we interpret the terms of reference of the probability assignable to a guess on this understanding. Seemingly, Bayes did recognise a nicety which is at the very core of what is most controversial at the present time, viz. the distinction between statements about the probability of events and statements about the probability of making correct assertions about events. Whatever we may justly say in criticism of Bayes, this much is certain. He never explicitly locates probability in the nebulous domain of the *mind*.

So far, there is nothing new in Section I of the memoir; and if we choose to verbalise (p. 117) the fifth proposition in terms of hypotheses rather than events, no confusion need arise when there is a factually realisable event corresponding to each hypothesis involved in the specification of a trial which involves *two* independent performances of randomwise sampling. One may indeed dismiss the likelihood that the statement of the *division* theorem would have led to any misunderstanding, were it not for the part it plays in Section II. The theme of the latter is the following model situation. Two balls successively come to rest on a smooth rectangular table at some point [*sic*]; and at this point we may discard the *ipsissima verba* of the author, if we wish to extract any intelligible message from the model. The truth is that this non-Euclidean and very real ball does not make contact with a Euclidean point. It comes to a standstill resting on a finite area of the surface of the said table. We may therefore convey the legitimate intentions of the author in the classical idiom, if we here interpret them against the background of a comparable, but factually realisable, situation unencumbered by Euclidean prepossessions.

Accordingly, we shall suppose that the table has equally spaced holes in rows and columns, the number of holes per row (i.e. number of columns) being  $a$ , numbered from the edge as 1, 2, 3 . . .  $a$ . If the number of holes per column (i.e. number of rows) is  $b$ , the total number of holes is  $n = ab$ . In accordance with Postulate 1 at the head of Section II of the memoir,\* we suppose that a ball released on the table will drop into any

\* Postulate 1. I suppose the square table or plane ABCD to be so made and levelled, that if either of the balls O or W be thrown upon it, there shall be the same probability that it rests upon any one equal part of the plane as another, and that it must necessarily come to rest.

single hole in the long run as often as it drops into any other, also that the ball will not rest till it falls into one of them. We may then write:

- $P_{kh} = n^{-1}$  the probability that a ball will drop into the  $k$ th hole of the  $h$ th row in a unit trial;  
 $P_h = b^{-1}$  the probability that a ball will drop into one or other of the  $a$  holes of the  $h$ th row in a unit trial;  
 $p_h = h \cdot b^{-1}$  the probability that a ball will drop into one of the  $h$  rows 1, 2 . . .  $h$  at a single trial.

We are now ready to undertake the 2-stage experiment which is the topic of Proposition VIII at the beginning of Section II. We take two balls A and B, release A first recording the row ( $h$ ) in which it comes to rest as the A-score, withdraw it and then release B successively  $r$  times recording  $x$  as the B-score of the  $r$ -fold trial if it comes to rest  $x$  times in one of the  $h$  rows 1, 2, 3 . . .  $h$ . The probability that the B-score will be  $x$  is accordingly

$$L_{x,hr} = r_{(x)} \cdot P_h^x \cdot (1 - p_h)^{r-x} = \frac{r_{(x)} \cdot h^x \cdot (b - h)^{r-x}}{b^r} \quad (\text{ii})$$

We now ask: what is the probability of the compound event that the A-score will be  $h$  and the B-score will be  $x$ ? By the elementary theorem of multiplication this is:

$$P_{hx} = P_h \cdot L_{x,hr} = \frac{r_{(x)} \cdot h^x \cdot (b - h)^{r-x}}{b^{r+1}} \quad . \quad . \quad (\text{iii})$$

Such is a fair translation of the factual content of Proposition VIII\*; and it registers no innovation. We shall now frame a different question, which is indeed novel. We postulate the following situation. We shall confine our attention to an infinite *sub-sequence* of trials in which the B-score is  $x$  as already defined. We shall then *guess* that the A-score in any such trial will be  $h$ . Stated in terms consistent with the Forward Look, the topic of Proposition IX is: *how often will our guess be right, if we consistently follow such a prescription for guessing?* In effect, the answer Bayes gives is the ratio of the infinite number of all

\* In the Euclidean continuum of Bayes the probability here defined by  $P_h$  will be infinitesimal. Bayes actually specifies a row-score lying between  $u$  and  $v$ ; and gives the result somewhat portentously as an area.



the probability of which nothing at all is known antecedently to any trials made or observed concerning it. And such an event, I shall call an unknown event. (*Italics inserted.*)

So far, we have kept close to the actual situation in which Bayes applies his theorem, deviating therefrom only in as much as we reject the concept of rest at a point as logically irrelevant and factually inappropriate. Factually, the rule given in Proposition IX of the memoir refers to a situation in which every possible outcome of the first two events is *equally probable* whence we may interpret with equal propriety either (iv) or (v) as the translation in modern symbolism of the theorem embodied in Proposition IX. Here we may break off to clarify the foregoing argument by recourse to a model situation more appropriate, if the end in view is to exhibit the factual implications of (iv) and (v) respectively. Our new model will be a box which contains six *unbiased* tetrahedral dice, which we shall specify as follows:

- (a) three such dice (*A*) having on one face 1 pip, on each of two faces 2 pips and 3 pips on the fourth face;
- (b) one such die (*B*) having on the four faces 1, 2, 3 and 4 pips respectively;
- (c) two such dice (*C*) having on each of three faces 2 pips and 1 pip on the fourth.

To make all the relevant information explicit, we may set out the data as follows, specifying score values as  $x$  and probabilities as  $y$ :

Type of Die	No. of Dice	Unit sample distribution
A	3	$x = 1 \ 2 \ 3 \ 4$ $y = \frac{1}{4} \ \frac{1}{2} \ \frac{1}{4} \ 0$
B	1	$x = 1 \ 2 \ 3 \ 4$ $y = \frac{1}{4} \ \frac{1}{4} \ \frac{1}{4} \ \frac{1}{4}$
C	2	$x = 1 \ 2 \ 3 \ 4$ $y = \frac{1}{4} \ \frac{3}{4} \ 0 \ 0$

# THE BACKWARD LOOK AND THE BALANCE SHEET OF THOMAS BAYES

By recourse to a chessboard lay-out we may derive the 2-fold toss score-sum ( $s_2$ ) distributions thus :

Type of Die	2	3	4	5	6	7	8
A	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	0	0
B	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
C	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	0	0	0	0

We may classify the outcome of the 2-fold toss in several ways of which one will suffice for our purpose, viz.: (i)  $s_2 < 4$ ; (ii)  $s_2 \geq 4$ . The corresponding probabilities for the three types of dice are then :

Die	(i) $s_2 < 4$	(ii) $s_2 \geq 4$
A	$L_{i.a} = \frac{5}{16}$	$L_{ii.a} = \frac{11}{16}$
B	$L_{i.b} = \frac{3}{16}$	$L_{ii.b} = \frac{13}{16}$
C	$L_{i.c} = \frac{7}{16}$	$L_{ii.c} = \frac{9}{16}$

We now define a *trial* in the following terms. A blind umpire shakes the box thoroughly to ensure random choice, selects a single die for a player who tosses it *twice*, records the joint score and replaces the die in the box. We may conceive an indefinitely large number of trials conducted in this way, the implications of the prescription being that the player's chance of getting a die of a particular type at a single trial is specifiable in terms of their relative numbers as follows :

TABLE I

Type	Chance of getting same
A	$P_a = \frac{1}{2}$
B	$P_b = \frac{1}{6}$
C	$P_c = \frac{1}{3}$

The entries  $P_a$ ,  $P_b$ ,  $P_c$  so defined are the so-called *prior probabilities* of the Bayes balance sheet. By the elementary rules of probability we can now set out as below the joint probabilities of: (a) getting a score  $s_2 < 4$  and  $s_2 \geq 4$ ; (b) choosing a die of one or other type:

TABLE II

Die	$s_2 < 4$	$s_2 \geq 4$	Total
A	$\frac{1}{2} \cdot \frac{5}{16} = \frac{5}{32}$	$\frac{1}{2} \cdot \frac{11}{16} = \frac{11}{32}$	$\frac{1}{2}$
B	$\frac{1}{6} \cdot \frac{3}{16} = \frac{1}{32}$	$\frac{1}{6} \cdot \frac{13}{16} = \frac{13}{96}$	$\frac{1}{6}$
C	$\frac{1}{3} \cdot \frac{7}{16} = \frac{7}{48}$	$\frac{1}{3} \cdot \frac{9}{16} = \frac{3}{16}$	$\frac{1}{3}$
Total	$\frac{1}{3}$	$\frac{2}{3}$	1

We have now classified the outcome of all trials which result in a score  $s_2 < 4$  under three headings w.r.t. the die selected by the player, viz.:

- (a)  $s_2 < 4$  with die of type A;
- (b)  $s_2 < 4$  with die of type B;
- (c)  $s_2 < 4$  with die of type C.

In conformity with the prescription of the game the long run frequencies associated with the three events are *in the same ratio* as the probabilities in the left-hand column of Table II. Their sum is  $\frac{1}{3}$ . If we confine our attention to the class of results defined by  $s_2 < 4$ , we may therefore assert that the long-run proportionate frequencies of such a score referable to each of the three types of dice are:

$$\begin{array}{ll}
 \text{A} & \frac{5}{32} \div \frac{1}{3} = \frac{15}{32} \\
 \text{B} & \frac{1}{32} \div \frac{1}{3} = \frac{3}{32} \\
 \text{C} & \frac{7}{48} \div \frac{1}{3} = \frac{14}{48} \\
 \text{Total} & 1
 \end{array}$$

In the phraseology of Bayes's successors, the adjusted frequencies in the right-hand column are the *posterior probabilities* of our balance sheet. In more general terms, they represent the long-run proportionate frequencies of samples whose source is one or other die in the class of all samples whose score has a

particular value. Symbolically, we may label the items of the final balance sheet as follows:

- $P_{h,r}$  the *prior* probability that any  $r$ -fold sample will come from a die of class H.  
 $L_{x,hr}$  the *conditional* probability that the  $r$ -fold sample score will be  $x$ , if the die is of class H.  
 $L_{h,xr}$  the *conditional* (so-called *posterior*) probability that the  $r$ -fold sample whose score is  $x$  will come from a die of class H.

By the elementary theorem of multiplication,  $P_{h,r} \cdot L_{x,hr}$  is the joint probability that the sample score will be  $x$  and that the source will be a die of class H. By the elementary theorem of addition, the unconditional probability that the sample score will be  $x$  is the sum of such products for all values of  $h$ . The law of the balance sheet is then in agreement with (iv):

$$L_{h,xr} = \frac{P_{h,r} \cdot L_{x,hr}}{\sum_{s=1}^{s=\infty} P_{s,r} \cdot L_{x,sr}} \quad . \quad . \quad . \quad . \quad (vi)$$

Let us now suppose that we decide to act in accordance with the following rule: whenever the score of the  $r$ -fold sample is  $x$  we shall say that the source of the sample is a die of class H; but we shall *reserve judgment* if the  $r$ -fold trial score is not  $x$ . In an indefinitely protracted sequence of trials, the proportion of samples referable to a die of type H among all samples whose score is  $x$  is  $L_{h,r}$ . Thus  $L_{h,r}$  is the long-run proportion of correct statements we shall make if we consistently follow the prescribed rule. If we specify as hypothesis H the particular assertion that the die is of class H, we may then say that  $L_{h,r}$  is the probability of choosing hypothesis H correctly by pursuing a prescribed course of behaviour regardless of the outcome of any single trial. Most assuredly, this is not the same thing as saying that  $L_{h,r}$  is the probability that hypothesis H is true if the score at a particular trial happens to be  $x$ . Likewise, and most assuredly, Bayes never employs this idiom, subsequently accredited to him by implication, in the exposition of Prop. IX in his memoir.

In the foregoing set-up we specify as hypothesis A the

assertion that the die cast is of type A. We may then choose to make the following rule: say that hypothesis A is true *whenever* the score  $x = s_2$  of the 2-fold toss is less than 4. Our balance sheet then shows that  $\frac{1}{3}\frac{5}{2}$  is the proportion of trials in which we select the die A if we confine our verdicts to the infinite class of trials which yield score values  $x < 4$  and reserve judgment about the die chosen in the infinite class of trials which yield score values  $x \geq 4$ . Thus  $\frac{1}{3}\frac{5}{2}$  is the probability of correctly choosing hypothesis A if we follow the prescribed rule consistently and  $\frac{1}{3}\frac{7}{2}$  is the probability of choosing it wrongly.

Alternatively, among a great number of conceivable rules of this sort, we may choose the following: say hypothesis B (i.e. the die is of class B) is true whenever the score of the 2-fold toss is at least 4. Within this infinite class of trials the proportion of trials in which we select the die B is

$$\frac{1}{6} \cdot \frac{1}{16} \div \left( \frac{1}{2} \cdot \frac{1}{16} + \frac{1}{6} \cdot \frac{1}{16} + \frac{1}{3} \cdot \frac{9}{16} \right) = \frac{1}{64} \simeq 0.20$$

In this case, the risk of falsely saying that the die cast is a die of type B is thus approximately  $(1 - 0.20)$  or 80 per cent. The reader will thus see what the so-called *posterior* probability of the event in the Bayes Balance Sheet means in terms of the classical theory of risk. We may speak of  $P_f = (1 - L_{h,x})$  as the risk of falsely choosing hypothesis H as the true one, if we follow the rule of saying that it is specified in advance. We may likewise call  $P_t = L_{h,x}$  the *stochastic credibility* of such a rule; and  $P_f = (1 - P_t)$  the *uncertainty safeguard* of the rule; but we shall deviate from the classical and behaviouristic approach referable to observable frequencies of external occurrences if we conceive  $L_{h,x}$  as the stochastic credibility of the hypothesis H or if we use a form of words assigning a probability  $P_f = (1 - L_{h,x})$  to its falsity.

The last model illustrates the implications of what many writers refer to as Bayes's theorem in its most general form, that is to say (iv) above. We could bring it into line with the actual model of Bayes by putting 9 dice, three of each type, in the box. The so-called prior probabilities ( $P_a = P_b = P_c = \frac{1}{3}$ ) are then equal. This imposes on the model the particular restriction embodied in the first postulate of Section II of the



*Essay*, and the equivalent but more general principle put forward in the Scholium. The reader may adjust the balance sheet accordingly. The expression for  $L_{h,x}$  now reduces to (v) which suffices for the Bayes model. For the class of 2-fold samples defined by  $x < 4$  we now have

$$L_{x,a} = \frac{5}{16} ; L_{x,b} = \frac{3}{16} ; L_{x,c} = \frac{7}{16}$$

$$\sum_{s=1}^{s=\infty} L_{x,s} = \frac{5 + 3 + 7}{16} = \frac{15}{16}$$

$$\therefore L_{h,a} = \frac{1}{3} ; L_{h,b} = \frac{1}{5} ; L_{h,c} = \frac{7}{15}$$

Let us now make the rule: say Hypothesis A is true when  $x < 4$ , otherwise say nothing. Our uncertainty safeguard for the rule so stated is now  $\frac{2}{3}$ , instead of  $\frac{17}{32}$  as for the preceding set-up.

Models of the type so far discussed are remote from the world's work. To get their relevance—if any—to practical affairs into focus, let us now see how we can bring this type of decision into workmanlike relationship to laboratory experience. So we now suppose that a laboratory stock of 1,000 female rats consists of: (a) healthy females which carry a sex-linked lethal gene; (b) normal healthy females. Accordingly, the probabilities that the individual offspring will itself be female are

$$p_a = \frac{2}{3} \text{ if the mother is a carrier}$$

$$p_b = \frac{1}{2} \text{ if the mother is normal}$$

For heuristic purposes, we may assume with sufficient plausibility that the elimination of male progeny by the lethal gene does not appreciably affect the total number of live-born rats, since there is commonly in rodents a considerable mortality of embryos owing to production in excess of uterine capacity. Let us then ask what long-run proportion of mothers of eight offspring, all female, are respectively carriers and

normal. In the foregoing symbolism we state the probability of each event on the appropriate assumption as:

$$L_{s,a} = \left(\frac{2}{3}\right)^8 \simeq 0.039 ; L_{s,b} = \left(\frac{1}{2}\right)^8 \simeq 0.0039$$

If we suppose that 5 per cent of the females are carriers, the prior probabilities are:

$$P_a = 0.05 ; P_b = 0.95$$

In accordance with (vi) we can thus state as follows the so-called posterior probability of the two events, viz. that the mother of eight all female is a carrier and that the mother of eight all female is normal as

$$L_{a,8} \simeq \frac{(0.05)(0.039)}{(0.05)(0.039) + (0.95)(0.0039)} \simeq 0.345$$

$$L_{b,8} \simeq \frac{(0.95)(0.0039)}{(0.05)(0.039) + (0.95)(0.0039)} \simeq 0.655$$

Of rat colonies with the prescribed composition, we may thus say elliptically: the probability of correctly asserting that a mother of eight all female is a carrier is  $P_i \simeq 0.345$ ; and the probability that such an assertion is false is  $P_j = (1 - P_i) \simeq 0.655$ ; but this form of words is legitimate only in so far as we assign the probabilities stated to the operation of the rule and then only if we suppose that we consistently follow some *randomisation recipe* for picking the rat-mother out of the colony. Moreover, we are not legitimately assigning a probability to an *individual verdict*. On that understanding, our judgments will be nearly twice as often wrong as right, if we do consistently follow the rule stated. Alternatively, we may decide to deem a female rat to be normal in the same set-up, if it has eight offspring all female. If so, our uncertainty safeguard will be 0.345 and the stochastic credibility of our assertion will be 0.655. In the long run we shall be about twice as often right as wrong.

Such a specification of so-called *posterior probabilities* in the current idiom of a calculus of judgments presupposes that we know the actual values of the prior probabilities  $P_a$  and  $P_b$ . In

many comparable situations we might know that the colony contained female rats of each of these two classes *alone* without knowing how many of each. Can our Bayes' balance sheet then lead us to formulate a rule with an assignable risk of error? Price thought, and Laplace thought, that the words italicised in the foregoing citation (p. 121) from the notorious Scholium suggest a way out of the difficulty, though it is not wholly clear that Bayes himself confidently proffered the suggestion for model situations to which the initial postulate of Section I has no relevance. The recipe known as the *Bayes' Postulate* is that we here assign equal probability ( $P_a = \frac{1}{2} = P_b$ ) to the choice of a sample from one or other class of rats when we are wholly ignorant of the true values. If we apply this convention to the foregoing problem, we derive

$$P_{a.s} = \frac{\frac{1}{2}(0.039)}{\frac{1}{2}(0.039) + \frac{1}{2}(0.0039)} \simeq 0.91$$

$$P_{b.s} = \frac{\frac{1}{2}(0.0039)}{\frac{1}{2}(0.039) + \frac{1}{2}(0.0039)} \simeq 0.09$$

The application of the Bayes' postulate thus confers the value 0.09 as the long-run frequency of false assertion referable to consistent application of the first rule stated above; but we know that the long-run frequency of false assertion in this situation is actually 0.655. Reliance on the Scholium here leads to a result inconsistent with the classical theory. Nor is this surprising, since it entails the vulgar error of *neglecting the population at risk*.

Our difficulties do not end here, if we assume, as do so many writers on statistical theory, that Bayes's theorem in one form or another embraces the terms of reference of a calculus of judgments. That we can rarely give numerical flesh and blood to the algebraic dry bones of the prior probabilities in a set-up of the sort last *discussed* is only one horn of the dilemma with which the invocation of the Scholium confronts us. There is another. We have here presumed the knowledge that the colony contains rat mothers of two sorts; but if we adopt the Scholium as a universal principle of statistical inference, we

have to provide for four situations in which the problem of identification may arise:

- (a) we both know that the colony contains two sorts of rats, and how many there are of each;
- (b) we know that the colony contains rats of both sorts, but we do not know how many there are of each;
- (c) for anything we do know to the contrary, the colony consists of one sort only;
- (d) we know that it does consist of one sort only, but we do not know which.

From the viewpoint of the mathematician we may say that the statement  $P_a = (1 - P_b)$  subsumes all four situations, if we allow  $P_a$  to have every value from 0 to 1 inclusive; but this evades a highly relevant *factual* aspect of the situation. If we say that  $0 < P_a < 1$  as in (a) and (b), we postulate a system of *randomwise* sampling in *two stages* in conformity with the model situation to which the balance sheet is factually relevant. If we say that  $P_a = 0$  or  $P_a = 1$  in conformity with (d), we say in effect that sampling occurs in *one stage*, in which event the Balance Sheet is irrelevant to the formulation of any rule of procedure consistent with the Forward Look, and the allocation of equal prior probabilities, i.e. the assumption of equal populations at risk, is inconsistent with what we already know about the situation.

Such inconsistencies will emerge more clearly when we have examined the model situation which led Laplace to break decisively with the classical tradition. Meanwhile, it may help us to sidestep some sources of confusion in contemporary controversy, if we here attempt a judicious assessment of what Bayes actually attempted to do without prejudging his intentions unduly. We may accordingly summarise the foregoing examination of his memoir as follows:

- (i) The theorem especially associated with the name of Thomas Bayes deals with *two* randomwise sampling processes, the outcome of *only one* of which is open to inspection.
- (ii) In the model situation to which the theorem expli-

citly refers any possible outcome of the first sampling process is as probable as any other.

(iii) For such a model situation, the result Bayes gives is entirely consistent with the classical theory of risks in games of chance, and the *ipsissima verba* of the author, if at times obscure, are not inconsistent with what we have elsewhere designated the *Forward Look*.

(iv) Bayes admittedly, but with some hesitation, advances the suggestion that we may assign equal probabilities to each possible outcome of the first sampling process when we have no certain information to the contrary; but he does so on the implicit assumption that the first sampling process is referable to *external occurrences*.

(v) We must allocate to Dr. Price alone the credit or discredit for extending the principle of insufficient reason to situations in which only one outcome of the first sampling process is realisable and to Price alone the responsibility for identifying the latter with a *subjective choice* in favour of one or other among different conceptual possibilities.

(vi) By this extension and identification, Price raised two issues which do not emerge in the model situation dealt with by the author:

(a) whether the choice between conceptually admissible hypotheses, only one of which can be applicable to a given situation, is factually on all fours with the concept of choice in the setting of the classical theory;

(b) whether there is any intelligible sense in which we can speak of such a process of sifting hypotheses as *random*, and as such relevant to the proper terms of reference of a stochastic calculus.

Of Bayes's suggestion, which invokes the principle of insufficient reason when the prior probabilities are severally referable to existential populations at risk, we may say that it is certainly reliable only when the populations at risk are equal; and there would be no need to invoke the principle if we had any means of knowing that this is so. Of his editor's implicit assumption that mental decisions made in total ignorance are random occurrences, it suffices to say that no one has hitherto

been able to advance any conclusive evidence in favour of a contention which is not amenable to experimental enquiry; but if we cannot plan an experiment to justify or to discredit it, we can at least devise experiments to test the validity of a more restricted contingent proposition, viz. do we actually make decisions of one sort or another with equal frequency when nothing we know about a situation furnishes any rational grounds for preferring one verdict to another?

Experiments of the sort referred to last are indeed easy enough to plan or to conduct. For example, we may: (*a*) tell each of a large group of people that the number of balls in a box is at least one and less than ten; (*b*) ask each member of the group to guess how many balls the box contains; (*c*) record the frequency with which individuals guess each of the numbers from 1 to 9 inclusive. To the writer's knowledge, no one has recorded a large-scale experiment of this type, perhaps because experience has taught us that threes, fives and sevens would turn up too often.

## CHAPTER SIX

### THE PLACE OF LAPLACE AND THE GRAND CLIMACTERIC OF THE CLASSICAL TRADITION

NOTABLE CONTRIBUTORS to knowledge in any age suffer from a double disability at the hands of history. Hero worshippers read into their authorities intentions that the authors themselves would not have dreamed of; and this is not necessarily flattering, since there is no one to one relation between hero-worship and intellectual good taste. Contrariwise, a chance remark wrested from its context for controversial advantage becomes the trademark of a philosophical system which critics can attack from the advantage of safe distance; and this is a rude sort of justice since contra-suggestible critics are not unduly willing to undertake a charitable appraisal of an opponent's viewpoint. So was it with David Hume. So may it well be with Bayes who did not explicitly state the theorem we associate with his name exhibited either as in (iv) of Chapter Four or in the particular form (v) which embodies the highly debatable axiom (p. 121) in the Scholium. The *fons et origo* of inverse probability is Laplace. For good or ill, the ideas commonly identified with the name of Bayes are largely his.

A demonstration, extensively discussed by Karl Pearson in the *Grammar of Science*, signalises both the explicit formulation of what we now choose to call Bayes's theorem and the use of the Scholium to validate judgments about the course of events. It appears in the *Memoire sur la Probabilité des causes par les événements*, "remarkable in the history of the subject," says Todhunter, "as being the first which distinctly enunciated the principle for estimating the probabilities of the causes by which an observed event may have been produced." The problem which Laplace there proposes is as follows. A player takes *with replacement*  $r$  balls\* from an urn known to contain white balls

\* Laplace actually speaks of *billets*, presumably thinking of the Royal lotteries referred to in Chapter Four. This may be the explanation of the vogue the urn model enjoyed.

and/or black ones. He observes that  $x$  of the  $r$  balls are white and  $(r-x)$  are black. According to Laplace, he may then conclude that the chance  $P_{(x+1),r}$  of getting a white ball at the  $(r+1)$ th draw is

$$P_{(x+1),r} = \frac{x+1}{r+2} \quad . \quad . \quad . \quad . \quad . \quad (i)$$

It will be easier to get into focus what are exceptionable assumptions in the derivation of this rule by Laplace himself, if we now construct a model which admissibly does lead to the solution he offers. We suppose that we have  $(n+1)$  urns each containing  $n$  balls, of which 0, 1, 2, 3 . . .  $n$  are white. Thus the proportions of white balls are 0,  $n^{-1}$ ,  $2n^{-1}$  . . . 1; and in general for the  $k$ th urn the probability of drawing a white ball in a single trial is  $p_k = kn^{-1}$ . In this set-up consecutive values of  $p_k$  differ by the same increment, viz.:

$$\Delta p_k = n^{-1}$$

Since there are  $(n+1)$  urns in all, the prior probability ( $P_k$ ) of choosing any one urn randomwise is  $(n+1)^{-1}$  for all values of  $k$ . If  $n$  is large we may therefore write

$$P_k \simeq \Delta p_k$$

We shall now denote by  $L_{k,x}$  what we may loosely and provisionally, but advisedly (*see* p. 143) here designate in the language of Laplace as the probability that the urn contains  $k$  white balls if the  $r$ -fold sample score is  $x$  white balls. Bayes's theorem gives:

$$L_{k,x} = \frac{P_k \cdot L_{x,k}}{\sum_{k=0} P_k \cdot L_{x,k}} \simeq \frac{L_{x,k} \cdot \Delta p_k}{\sum_{k=0} L_{x,k} \cdot \Delta p_k}$$

In this ratio:

$$L_{x,k} = r_{(x)} \cdot p_k^x (1 - p_k)^{r-x}$$

In the last expression  $r_{(x)}$  is a constant, since we are talking about a specified *particular* sample size and a specified score.



If the sample is finite and we sample with replacement,  $n$  being in any case very large in comparison with  $r$ :

$$\sum_{k=0}^{k=n} r_{(x)} \cdot p_k^x (1 - p_k)^{r-x} \cdot \Delta p_k \simeq r_{(x)} \int_{p=0}^{p=1} p^x (1 - p)^{r-x} \cdot dp$$

$$\therefore L_{k,x} \simeq \frac{p_k^x (1 - p_k)^{r-x} \cdot \Delta p_k}{\int_0^1 p^x (1 - p)^{r-x} \cdot dp}$$

The conditional probability of getting a white ball at the  $(r + 1)$ th trial is  $p_k$ , if the urn is the  $k$ th. The conditional probability that the urn contains  $k$  white balls if the  $r$ -fold sample score is  $x$  is  $L_{k,x}$  as above. In conformity with our assumption that  $n$  is very large, the conditional probability of getting a white ball from an urn containing  $k$  white balls at the next trial after getting  $x$  white balls in  $r$  trials from the same urn is therefore:

$$P_{(x+1),kr} = L_{k,x} \cdot p_k \simeq \frac{p_k^{x+1} (1 - p_k)^{r-x} \cdot \Delta p_k}{\int_0^1 p^x (1 - p)^{r-x} \cdot dp}$$

The unconditional probability of getting a white ball at the  $(r + 1)$ th trial follows from the addition theorem, viz.:

$$\begin{aligned} P_{(x+1),r} &= \sum_{k=0}^{k=n} P_{(x+1),kr} \\ &\simeq \frac{\int_0^1 p^{x+1} (1 - p)^{r-x} \cdot dp}{\int_0^1 p^x (1 - p)^{r-x} \cdot dp} \\ &\simeq \frac{B(x + 2, r - x + 1)}{B(x + 1, r - x + 1)} \end{aligned}$$

The ratio of the two Beta functions above is:

$$\frac{\Gamma(x+2) \cdot \Gamma(r-x+1)}{\Gamma(r+3)} \cdot \frac{\Gamma(r+2)}{\Gamma(x+1) \cdot \Gamma(r-x+1)}$$

$$= \frac{(x+1)! (r+1)!}{(r+2)! x!} = \frac{x+1}{r+2} \quad (\text{ii})$$

So far the algebra. We have still to divulge what this charity is in aid of, and if the reader is perplexed, it is not without good reason. Equation (ii) is identical with (i) above, but the problem of which it is the solution is not the problem Laplace himself sets out to solve. The latter is a problem about *one urn and only one*. Its statement admits no preliminary random choice of the urn from which we draw the  $(r+1)$  balls. It is therefore not a situation which we can discuss in terms to which the balance sheet of Bayes is factually relevant, unless we know everything relevant about its contents, in which event we already know the solution.

This confusion of topic would scarcely call for comment, if Laplace had offered the solution of the problem as a *jeu d'esprit* to entertain the wits of a Paris salon. It is, it happens, the foundation-stone of a new system of thought. The urn from which we are to draw the white and black balls is no less than the one to which Quetelet refers when he epitomises a novel interpretation of scientific reasoning in the assertion "*l'urne que nous interrogeons c'est la nature*." Pearson speaks of it (*Grammar of Science*, 2nd edn., p. 147) as the *nature bag*. Among other consolations he extracted therefrom, Laplace conceived it as possible to provide a long overdue rationale for inductive reasoning and a rebuttal of Hume's so often misconstrued scepticism. Thus he uses it to solve to his own satisfaction a problem cited in the appendix Price wrote to the *Essay* of Bayes as a proper application of the Scholium. Can we assign a probability to the assertion that the sun will rise tomorrow? On the understanding that it has done so daily for 5,000 years (1,826,213 days), the formula of Laplace assigns  $P_{r+1} = \frac{1826214}{1826215}$  or betting odds of 1826214 : 1 in favour of the event.

In the context of contemporary discussion, this use of his theorem has a lighter side with which Karl Pearson deals earnestly and at length. Following Eddington, recent writers on scientific method have welcomed the reformulation of scientific laws in statistical terms which supposedly dispense with the eminently operational concept of causality and the shadow of an absolute, if none the less unattainable, truth in the background. The implication is that miracles can happen, and hence that a long-standing antagonism between the laboratory and the vestry admits of a happy ending.

Unhappily, Laplace is no longer able to comment on this engaging prognosis of his labours; and Pearson's major pre-occupation with the exposition of his views when he speaks of the nature bag will not help the zealous seeker after truth to find in (i) above the spiritual solace which Eddington wistfully proffers. His ensuing remarks (*op. cit.*, pp. 142-3) are therefore worthy of reproduction:

Laplace has even enabled us to take account of possible "miracles," anomalies, or breaches of routine in the sequence of perceptions. He tells us that if an event has happened  $p$  times and failed  $q$  times, then the probability that it will happen the next time is  $\frac{p+1}{p+q+2}$ , or the odds in favour of its happening are  $p+1$  to  $q+1$ . Now if we are as generous as we possibly can be to the reporters of the miraculous, we can hardly assert that a well-authenticated breach of the routine of perceptions has happened *once* in past experience for every 1,000 million cases of routine. In other words, we must take  $p$  equal to 1,000 million times  $q$ , or the odds against a miracle happening in the next sequence of perceptions would be about 1,000 millions to one. It is clear from this that any belief that the miraculous will occur in our immediate experience cannot possibly form a factor in the conduct of practical life. Indeed the odds against a miracle occurring are so great, the percentage of permanently diseased or temporarily disordered perceptive faculties so large as compared with the percentage of asserted breaches of routine, and the advantage to mankind of evolving an absolutely certain basis of knowledge so great, that we are justified in saying that miracles have been *proved* incredible—the word *proved* being used in the sense in which alone it has meaning when applied to the field of perceptions.

The system of thought which Laplace thus introduced departs from the pre-existing tradition explicitly at three levels; and the theorem under discussion illustrates them all. To the writer's knowledge only one exponent of the theory of probability has explicitly pointed out all the debatable issues involved. In *Probability, Statistics and Truth*, p. 175, R. von Mises writes as follows:

Nearly all the older textbooks were agreed that these problems involve a special kind of probability. They called it the *probability of causes*, or the *probability of hypotheses*, and assumed it to be different from the usual probability of events. The word "cause" can, in fact, be easily introduced into the description of the Bayes' problem. We just say that the "cause" of result  $n_1$  6's in  $n$  trials is a special value of  $x$ , where  $x$  is the probability for a given die showing a 6 when thrown. Since we wish to know the probability of different values of  $x$ , we can say that we are examining the probabilities of different causes. This is nothing but an idle play upon words. We actually calculate the probability of a certain event in a certain collective. . . . Its elements are *experiments carried out in two stages*; drawing a stone from an urn, and throwing it out  $n$  times from a dice box. The probability is defined in this case, exactly as in all the others, as the limiting value of the relative frequency. No excuse can be given for speaking of a special kind of *probability of causes* or *probability of hypotheses*, since all that is determined is, as always, the probability of an event in a collective.

In what follows, I shall deal separately with the debatable issues on which von Mises puts the spotlight in the preceding citation, viz.:

- (a) in what class of situations is the concept of *prior* probability factually relevant;
- (b) in what sense, if any, can we legitimately identify the concept of *posterior* probability with the probability of *causes*;
- (c) what meaningful content can we convey by the term probability of *hypotheses* within the framework of a behaviouristic attitude?

*The Concept of Prior Probability.* We have seen that the actual model which leads to the solution given by (i) is an infinitude of urns to which we assign different relevant parameter values

$p_k$  equally spaced and all different. Thus it is a property of our stratified universe of urns that the distribution of prior probabilities is both rectangular and continuous. The introduction of the assumption of continuity in this context is not without interest, because the differential calculus had no function in the classical theory except as a convenient computing device to sidestep laborious calculation in the domain of finite differences. Its utility so conceived is purely empirical; and its legitimate use presupposes *ad hoc* investigation to justify the assumption that the order of error in the appropriate summation of relevant terms of a discrete probability series does not contravene the operational intention. Henceforward, theoretical distributions of this or allied types, here the so-called incomplete beta function (Type II of Pearson's system), acquire an indispensable theoretical status in virtue of the principle which prompted Laplace to formulate a distribution of prior probabilities as stated. The distribution of  $p$  must be continuous to allow for the possibility that the unknown  $p_k$  may have any value if we start with no background information about the model set-up. Likewise, it must be rectangular to accommodate the postulate that all prior probabilities are equal when we know nothing relevant about the source of a sample other than its definitive score.

The two issues last stated invite separate consideration. The second alone has been, and still is, the pivot of a controversy which might be more profitable if it took within its scope the first. In assigning equal prior probabilities to the sources of the sample when we start with no background knowledge, Laplace takes over what is most exceptionable in Bayes's Scholium, henceforth referred to as the *principle of insufficient reason*. He proffers it as a new law of thought, commonly accepted till the time of Boole, but subsequently contested for different reasons by Boole's successors. The usual arguments advanced against the principle of insufficient reason as a self-evident axiom of stochastic inference in the domain of factual research are:

- (a) that we do not commonly investigate the credentials of a hypothesis unless we either have good initial reasons for

supposing that it is faulty or have good initial reasons to commend its claims\*;

(b) even if the investigator undertakes an enquiry with an *open mind*, i.e. a mind uninformed by relevant knowledge of his materials in contradistinction to a mind disciplined to accept the verdict of ineluctable new fact, there is no particular reason to assume that his or her mental valuations have any relation to the complexities of natural phenomena.

Few hard-headed investigators will contest the first contention; but the second concedes to the opposition an unnecessary advantage which will be easy to recognise, if we now focus attention on a type of problem which has lately become a playground for stochastic test and estimation procedures, viz. the therapeutic trial. If one asks whether treatment B is more efficacious than treatment A, the orthodox perception sets out in effect to explore three *conceptual* possibilities:

- (i) B is in fact better than A;
- (ii) A is in fact better than B;
- (iii) A and B are equally good.

In this context, *good* is referable to a population and has no relevance to the individual as such; but this limitation need not concern us here. When we then say that the end in view is to arbitrate on the merits of these conceivably correct hypotheses, we imply that one, *and only one*, of these hypotheses is

\* F. J. Anscombe (1948) states it thus:

It is highly unusual for an experiment to be conducted without some prior knowledge of expectation of the result, and the case of complete ignorance should be merely a stepping-stone in the development of the theory. The special virtue of a theory of rational belief is surely that it can take account of even the vaguest prior knowledge. Some such knowledge generally exists even when a hypothesis is first propounded by an experimenter, before it has been tested; since some conceivable results of an experiment carried out to test it would be classed by him as "reasonable" or expected, and others as utterly unlikely. (*Journ. Roy. Stat. Soc.*, CXI, p. 193.) Elsewhere (*Mind*, Vol. 60) the same author rightly comments in the same vein as follows:

As soon as any proposition or hypothesis has been formulated which is worth testing experimentally, there is already evidence as to its truth derived from existing accepted knowledge and from considerations of analogy or "consilience." A question to which we have no grounds whatever for hazarding an answer is an idle question and would not be the subject of scientific investigation.

correct. It is equally implicit in the second objection to the Bayes' postulate, and in the viewpoint of those who subscribe to it, that a decision of this sort is on all fours with a decision about which of the three types of dice we have taken from the box in the model situation of p. 122. Now this assumption leads to a manifest inconsistency. If we look more closely at the problem of the therapeutic trial we may say that there exist in the box two sorts of dice which we label as A and B without knowing whether the fact that we do label them in this way has any relevance to the long-run mean score we get from tossing a die of either sort. The very fact that we admit as a topic of enquiry Hypothesis (iii) above, viz. that treatment A and treatment B are equally good, is also an admission of the possibility that our classification has no factual bearing on the long-run mean score of the  $r$ -fold toss, i.e. that all the dice in the box are of one and the same type in any sense relevant to the classical calculus of risk in games of chance.

In Damon Runyon's vivid idiom there is clearly no percentage in looking at the problem of the trial in this way. In the real world only one hypothesis, as here stated, can be right. If so, every type of die we conjure into our model to specify an additional and false hypothesis exists only in the eye of the beholder. By the same token, probability is a state of mind and prior probability is merely a device for registering a purely subjective evaluation. If so, we exclude the possibility of defining the proper terms of reference of a calculus of judgments unless we first embrace philosophical idealism; but we limit the scope of profitable discussion in this way merely because we have chosen the wrong classical model for the problem in hand. It is wrong for the reason stated by von Mises. The concept of prior probability is factually relevant only to *experiments carried out in two stages*.

Once we lose grip of this elementary consideration, as indeed we do if we invoke the concept of prior probability in the domain of the therapeutic trial, we depart from a behaviourist viewpoint, because we have ceased to communicate within the framework of observable events. In the Bayes' model each hypothesis to which we assign a finite prior probability is referable to an *existent* population at risk; and we audit the

balance sheet on the assumption that we are free to choose randomwise a die of a particular type in a box *before* tossing it or a particular urn or bag *before* drawing balls from it. In the set-up of the trial we have no such *preliminary act of random choice*. We have to take nature on her own terms.

That so simple a consideration does indeed merit emphasis will be evident if I cite a passage in which Karl Pearson expounds the Laplace theorem discussed above. There we have one bag with two sorts of balls, albeit in the original publication an urn with two sorts of tickets, but in what proportions we do not know. As the tale unfolds, this mysterious bag turns out to be an infinitude of bags, all except one of which exists only in the mind of the Maestro; but it miraculously regains its status as the one bag of the problem for which Laplace offers (i) as the correct solution. Thus Pearson says\*

We are now in a position to return to our bag of white and black balls, but we can no longer suppose an equal number of both kinds, or that routine and breach of routine are equally probable. We must assume our "nature bag" to have every possible constitution or every possible ratio of black to white balls to be equally likely; to do this we suppose an infinitely great number of balls in all. We may then calculate the probability that with each of these constitutions the observed result, say  $p$  white balls and  $q$  black balls (or,  $p$  cases of routine and  $q$  anomalies) would arise in  $p + q$  drawings.† This will determine, by Laplace's principle, the probability that each hypothetical constitution is the real constitution of the bag.

The reader will not be slow to detect that nothing about this surprisingly one and indivisible *nature-bag* remains constant throughout the syntactical operations of the previous paragraph except—and then only by implication—the material used in its manufacture before insertion of its appropriate but none the less mutable contents. Possibly also its shape remains the same, if we assume that the said material is fairly stiff. In that event the author of the *Grammar of Science* might have referred more felicitously, if more sepulchrally, to the original urn. In so far as it is relevant to the argument, all that we learn about our

\* *Grammar of Science*, 2nd edn., p. 147.

† The reader may suppose the ball returned to the bag after each drawing.



bag with its *two-fold* equipment of balls is that its contents change from sentence to sentence with a teasing disregard for the plural flexion of the English noun. So stupendous an exploit of stochastic conjuring should suffice to emphasise the importance of what von Mises means when he speaks of an *experiment carried out in two stages*. The absurdities into which the principle of insufficient reason leads us arise less from the circumstance that it is gratuitous than because it is wholly irrelevant to most situations in which even its opponents would readily concede its convenience if susceptible of proof. This will emerge more clearly when we examine the properties of a model situation discussed below.

*The Probability of Causes.* We have seen that contemporary writers who repudiate with abhorrence the doctrine of inverse probability commonly do so because the allocation of equal prior probabilities is uncongenial to the outlook of the laboratory worker with first-hand knowledge of his materials. However, they rarely do so with explicit recognition of the factual relevance of the concept itself to most situations which prompt the theoretician to condone its invocation. Still less do they reject the doctrine of inverse probability for reasons foreshadowed in the foregoing discussion of Bayes's *Essay*. Thus those who knowingly reject it, relinquish the principle of insufficient reason somewhat wistfully, seemingly unaware of the most insidious departure from the classical tradition implicit in the exposition of the doctrine by Laplace himself.

For expository purposes, let us now assume for the time being that our infinite collection of urns is indeed a model relevant to the identification of a correct hypothesis. Can we then say that  $L_{h,x}$  is the probability of a *cause* in the sense that we have loosely, provisionally and admittedly in the language of Laplace himself, specified it (p. 134) as the probability that the source is the  $k$ th urn if the score is  $x$ ? This form of statement is highly elliptical, and conveys something which is inconsistent with the Forward Look. It suggests that we first look at a sample, then make an assessment of a probability on the basis of the evidence it supplies, an implication inherent in the misleading designation *posterior* applied thereto. Now we can give no

meaning to the probability we are attempting to assess when we look at the issue in this way unless we abandon the factual historic framework of the classical theory of risks. Within that framework we must interpret  $P_f = (1 - L_{h,x})$  as the *uncertainty safeguard* of a rule of procedure stated in advance and consistently followed regardless of the outcome of any individual trial.

So conceived  $L_{h,x}$  is not the probability of a cause inferred from an isolated event. It is the probability of correctly identifying a cause, if we pursue throughout an endless succession of a particular class of events one and the same rule specified *in advance* by the occurrence of a particular score value. To me, it thus seems that so much emphasis on the credibility of the principle of insufficient reason has diverted attention from a highly exceptionable innovation which alone endorses the literal meaning of the epithet *inverse* in the doctrine of Laplace. If we concede to tradition by identifying the stochastic credibility of a rule with a *posterior* assessment of probability conceived in terms of *weighing the evidence* when we already know the sample score, we have taken with Laplace a decisive backward step. We have abandoned the Forward Look of the Founding Fathers.

*The Probability of a Hypothesis.* Having taken this step, our descent into a nebulous domain of mental images is swift. We can speak of the probability of correctly guessing that hypothesis K is the right one in strictly behaviourist terms, only if we confine our statements to the verbo-factual level at which we can assign a limiting ratio to the frequency of two classes of external events within a framework of consistent conduct. Our concern is then with: (a) statements we make to the effect that hypothesis K is correct whenever we encounter a score  $x$ ; (b) trials of which it is both true that hypothesis K correctly identifies the source of the sample and true that  $x$  is also the sample score. Needless to say, this presupposes a correct factual specification of our statistical model. If we invoke Bayes's theorem we then do so on the understanding that our sampling procedure is factually consistent with the choice of a sample with score  $x$  from a source specified by hypothesis K in the endless sequence of such trials.

When we speak of the probability that hypothesis  $K$  is true on data  $k$  in the idiom of Carnap and Jeffreys, we implicitly abandon the attempt to locate in the world of affairs this preliminary act of choice in a 2-stage experiment for which the balance sheet of Bayes is unexceptionable in terms of the classical theory of risks. The sources specified by our hypotheses—of which we may postulate, as does Laplace in the above treatment, an enumerable infinitude each one equally admissible—are no longer observable entities with respect to which each relevant prior probability is referable to an existent population at risk. They are now conceptual images concerning which the classical theory of risks is entirely silent. Accordingly, we then formulate Bayes's theorem in terms such as the following:

Suppose, in fact that an event can be explained on the mutually exclusive hypotheses  $q_1 \dots q_n$  and let  $H$  be the data known before the event happens, so that  $H$  is the basis on which we first judge the relative probabilities of the  $q$ 's. Now suppose the event to happen. Then Bayes' theorem states that the probability of  $q_r$  *after it has happened* (i.e. on data  $H$  and  $p$ ) varies as the probability *before it happened* multiplied by the probability that it happens on data  $q_r$  and  $H$ . The probability  $P(q_r/pH)$  is therefore called the *posterior* probability,  $P(q_r/H)$  the *prior* probability, and  $P(p/q_rH)$  will be called the *likelihood*. (Kendall's *Treatise*, Vol. I, p. 176.)

I do not cite this passage (*italics* inserted) to score a debating point by placing on record the difficulty of understanding what an event is unless it happens. Nor is it necessary to reiterate that the event which has or has not happened, as the case may be, is in this context two totally different events referable to different stages of a 2-stage experiment. I do so especially because I want to emphasise that some current writers who decline to endorse the principle of insufficient reason subscribe both: (a) to the misconception that the concept of prior probability is essential to the definition of the terms of reference of a calculus of judgments; (b) to the belief that the theory of probability endorses posterior judgments on the basis of isolated events. This misconception of the legitimate status

both of prior probability and of posterior probability in the classical theory of risks is indeed inherent in the symbolism adopted by Jeffreys, by Carnap and by other writers who more candidly disclaim the attempt to define probability in terms consistent with a behaviouristic outlook.

*A Generalised Model of Stochastic Inference.* We have seen that the formula (p. 134) which epitomises the announcement of the doctrine of inverse probability is not indeed a solution relevant to the properties of the factual model situation Laplace proposes to explore and to expound. We have also seen that his approach is inconsistent with a behaviouristic—or as a previous generation would have said, an objective—attitude towards the relevance of an axiomatic theory of probability to the real world. We shall be able to get the issues raised above in clearer perspective if we now examine, both from his own viewpoint and from one which is consistent with the behaviourist approach, a model situation mentioned by Laplace himself.

In the *Introduction a Théorie analytique des Probabilités* (1820) Laplace proposes the following problem. An urn contains two balls. Initially, one knows that both may be white, both may be black or that there may be one ball of each sort. A player draws out one ball randomwise, replaces it and draws again. If the ball is white at each draw, what is the probability that it will be white at the third after replacement of the ball extracted at the second?

Laplace argues as follows. One need here advance only two hypotheses: (a) neither of the balls is black; (b) one only is black. On hypothesis (a) the likelihood of drawing two white balls in a double draw is 1, on hypothesis (b) it is  $\frac{1}{4}$ . He regards it as both meaningful to assign prior probabilities to hypotheses as such and legitimate to equate prior probabilities when we have no information to discredit the assumption. Accordingly, he assigns to the truth of hypothesis (a), in accordance with (iv) in Chapter 5, the probability

$$\frac{\left(\frac{1}{2}\right) (1)}{\frac{1}{2} (1) + \frac{1}{2} \left(\frac{1}{4}\right)} = \frac{4}{5}$$

Similarly, he assigns to the truth of hypothesis (*b*)

$$\frac{\frac{1}{2} \left(\frac{1}{4}\right)}{\frac{1}{2} (1) + \frac{1}{2} \left(\frac{1}{4}\right)} = \frac{1}{5}$$

Laplace speaks of each such ratio as the probability of the *cause*. Having done so, he invokes a retrospective interpretation of the addition theorem in the following terms (7th Principle). At the third trial the probability of drawing a white ball will be 1 if hypothesis (*a*) is correct and  $\frac{1}{2}$  if hypothesis (*b*) is correct. The joint probability that (*a*) specifies the cause correctly and that the ball will be white is for Laplace  $\left(\frac{4}{5}\right) (1)$ . The alternative that (*b*) specifies the cause correctly and the ball will be white he cites by the same token as  $\left(\frac{1}{5}\right) \left(\frac{1}{2}\right)$ . Since one or other hypothesis must correctly specify the cause of the event, he derives the probability of getting a white ball as:

$$\left(\frac{4}{5}\right) (1) + \left(\frac{1}{5}\right) \left(\frac{1}{2}\right) = 0.9$$

In the idiom of Chapter Five, we may say that Laplace here assigns an uncertainty safeguard  $P_f = 0.1$  to the rule of consistently betting that the third ball is white, when the first two of a 3-fold trial are also white. Let us now look at the problem in terms consistent with the Forward Look. The results of two particular unit trials admittedly leave us with only two hypotheses, as stated; but the divulged result of one 2-fold trial will merely restrict the verbal form of the bets we may choose to make from three to two. If hypothesis (*a*) is true, the probability that the ball will be white in each of an unending sequence of single trials is unity. If it is false, the probability is  $\frac{1}{2}$ . All we know is that it is either true or false. Hence we can say that at least 50 per cent of our guesses will be right if we bet on white. Alternatively, the uncertainty safeguard of the rule is  $P_f \leq 0.5$ .

Let us next suppose that we can draw *six* balls successively and randomwise with replacement, equipped as before with the knowledge that one 2-fold trial yielded two white balls, and hence with only two relevant hypotheses. We are free to make the rule: say that hypothesis A (that neither of the balls

is black) is true when all six balls are white, otherwise reject it as false. Three situations then arise:

- (i) we rightly reject hypothesis A when there is at least one black ball in the sequence of six;
- (ii) we wrongly accept hypothesis A when all six balls are white;
- (iii) we rightly accept hypothesis A when all six balls are white.

By definition, we do not err if (i) holds good, and we do not err if (iii) holds good. The probability assignable to a run of 6 white if hypothesis (*b*) is true is  $2^{-6}$ , and in that event we wrongly accept hypothesis (*a*). Since (i) and (iii) entail no risk, our risk cannot be greater than  $2^{-6} = 0.015625$ , and we shall accordingly assign the uncertainty safeguard of our rule as  $P_f \leq 0.015625$ . More generally, we may state our rule in the form: reject hypothesis A if at least one black ball turns up in the  $r$ -fold trial and accept it if all the balls are white. We may then tabulate as follows the risk associated with trials of 1, 2, etc.:

$r$	$P_f$	$r$	$P_f$
1	0.500	4	0.062500
2	0.250	5	0.031250
3	0.125	6	0.015625

Should a player, who bets in conformity with the rule last stated, wish to guarantee himself against bankruptcy without incurring a risk greater than 5 per cent, he must therefore: (*a*) fix his stake and the maximum number of  $r$ -fold trials on which he will bet in accordance with his capital resources at the start; (*b*) get his opponent to agree *in advance* that the number ( $r$ ) of unit trials per game is not less than 5.

If we thus dispose of the problem as stated by Laplace in terms consistent with the classical theory for division of stakes, the only legitimate use of our knowledge about the outcome of an antecedent 2-fold trial is to exclude the hypothesis that both the balls in the urn are black, and limit accordingly the form our rule will take. The reader may therefore ask whether the

contingency we are discussing is not in fact reducible to the question: what is the probability that we shall pick three white balls in succession? In one sense this is so, but if so, we can incorporate the prescription of the outcome of the first two of the three unit trials in a rule *stated in advance* as above, only if we also take the risk of never being able to make a bet. In terms of the classical theory we then interpret the problem of the two balls as the problem of assigning the risk of betting that the third ball drawn in a 3-fold trial will be white, if: (a) the first two are also white; (b) our only initial information is that either ball may be either white or black. It remains true that the risk of erroneous statement is  $P_f \leq 0.125$ , if we bet consistently to that effect; but we have now conceded the possibility that both balls may be black. In that event our rule is useless, in so far as it deprives us of the right to make a bet on the outcome of *every* 3-fold trial.

To cover the contingency last stated we must extend the terms of reference of the rule. For instance, we may:

- (i) say that both the balls in the urn are white, whenever all three balls drawn in a 3-fold trial are white;
- (ii) say that both the balls in the urn are black, whenever all three balls drawn in the 3-fold trial are black;
- (iii) otherwise, say that one of the balls in the urn is white, the other black.

If both balls are black or if both are white, our rule will never lead us astray. If one is white and one is black, the probability of getting three white balls or three black balls is  $2(2^{-3}) = 0.25$ , in which event our verdict will be wrong. Whence the probability of error cannot exceed 0.25.

The inconsistency of the classical theory and the solution of Laplace emerges more clearly if we elaborate the problem he himself states, e.g. what is the probability that the next five balls we draw will be white, if the two balls previously drawn are white. The answer Laplace would give is:

$$\left(\frac{4}{5}\right) (1) + \frac{1}{5} \left(\frac{1}{32}\right) = \frac{129}{160} \simeq 0.8065$$

Thus the risk of betting accordingly is  $P_f \simeq 0.1935$ . The only

answer consistent with the classical approach is that the risk cannot be greater than  $2^{-7}$ . Whereas the classical theory thus assigns a risk  $P_f \leq 0.5$  on the outcome of betting on white at the third draw when the *Théorie Analytique* assigns  $P_f = 0.1$ , the classical theory assigns to betting on white at the third, fourth, fifth, sixth and seventh draw when the first and second are white a risk  $P_f \leq 0.0078125$  and the *Théorie Analytique* assigns  $P_f = 0.1935$ .

In the solution of this problem proffered by Laplace, three principles alien to the classical theory intrude. In terms of the classical theory, we must interpret Bayes's theorem as a rule of guessing with an assignable risk referable to an endless sequence of trials. Laplace interprets it as a *retrospective evaluation of the result of an individual trial divorced from the series of the event*. In terms of the classical theory Bayes's theorem refers to a two-stage sampling process in the domain of external events. In the use Laplace makes of it, the first stage is a mental artefact which has no reference to observable occurrences. In so far as it is meaningful in the context of the classical theory to speak of the probability that a particular hypothesis is true, i.e. that we shall win our bet that it is true, the only values assignable to the prior probabilities in this set-up are zero or unity. Laplace himself assigns to each of two conceptually admissible hypotheses a prior probability of 0.5 and to the third zero prior probability on the basis of information the appropriate betting rule cannot anticipate, if stated before the game begins.

Interpreted as in Chapter Five in terms consistent with the Forward Look, and formally embracing as a limiting case situations in which the first stage of the sampling process is non-existent, Bayes's theorem leads to no inconsistencies or paradoxes which we encounter in the *Théorie Analytique*. This will be clear if we now *frame a rule which takes within its scope situations to which Bayes's theorem is factually relevant, doubtfully relevant or irrelevant*. We shall be able to do so only if we content ourselves with defining the upper limit of the uncertainty safeguard.

We shall assume that a bag contains pennies which *may* be of three sorts only: (a) with the Queen's head on both sides; (b) with Britannia on both sides; (c) with the Queen's head on



one side and Britannia on the other. We shall further assume that (c) are "unbiased," i.e. that  $p = \frac{1}{2} = 1 - q$  is the probability of scoring a head at a single toss. The use of the word italicised in the first sentence is intentional to cover four of the possibilities w.r.t. our background knowledge of the situation:

- (i) we know both that the bag contains pennies of each sort and how many of each sort it contains;
- (ii) we know that the bag contains pennies of each sort but we do not know how many of each it contains;
- (iii) we do not know whether the bag contains pennies of all three sorts, two only or one only;
- (iv) we know that the bag contains pennies of only one sort but we do not know which sort.

Of these admissible situations (i) alone fulfils the condition to which Bayes's theorem, as defined by (v) in Chapter Five is unquestionably relevant in the domain of action, i.e. in the sense that all the relevant numerical data are at our disposal. The definition of (iv) excludes the factual possibility of sampling conceived as a two-stage process. If we conceive it *formally* in this way as a limiting case of (i), we can do so only if we say that one of the three prior probabilities must be unity, the other two being zero. This is inconsistent with the principle of insufficient reason, which would assign the numerical value  $0\cdot\dot{3}$  to each. Of the remainder, (ii) is the only one to which the Scholium has factual relevance if admissible; but in real life we shall rarely encounter situations in which we can distinguish (ii) from (iii) which embraces both (ii) and (iv) as limiting cases. Clearly, (iv) is inconsistent with the Scholium, and we must also regard (iii) as inconsistent therewith.

Thus we have before us four situations to only one of which the principle of insufficient reason is intelligibly relevant; but in real life we cannot say that this is necessarily true of the third. It may happen that we start with some knowledge of the source of the pennies in the bag, knowing also that errors of minting which give rise to the bad pennies of classes (a) and (b) are extremely rare. If so, our background knowledge, though inadequate to prescribe an exact figure for the corresponding

# STATISTICAL THEORY

prior probabilities  $P_a, P_b, P_c$ , may fully justify our conviction that  $P_c$  is much greater than  $0.3$ ; and this again is inconsistent with the principle of insufficient reason.

Let us now presume that an observer chooses randomwise at each trial one penny from the bag, tosses it  $r$  times, observes the head score  $x$  and records his conviction that it is a penny of one of the three specified types (a), (b), (c) above. We may then use a model bag from which we *may* extract (in the general sense defined) a penny belonging to one of these three classes to illustrate the statement of a comprehensive rule which usefully covers all the situations specified by (i)–(iv) above. We shall initially suppose that the bag contains 100 pennies as follows:

No. of Pennies	Type	Corresponding Hypothesis (H)	Prior Probability ( $P_h$ )
5	(a) with two heads	A	$P_a = 0.05$
20	(b) with two tails	B	$P_b = 0.20$
75	(c) normal unbiased	C	$P_c = 0.75$

We may set out as below the conditional probability (likelihood) for a score of 0, 1–3 and 4 heads in a 4-fold toss:

For Penny of Type	Conditional Probability of Score of		
	0	1–3	4
A	0	0	1
B	1	0	0
C	$\frac{1}{16} = 0.0625$	$\frac{7}{8} = 0.875$	$\frac{1}{16} = 0.0625$

The complete Bayes Balance Sheet for the 4-fold toss is therefore:

Type of Penny	Probability of a score of		
	0	1–3	4
A	$(0.05)(0) = 0$	$(0.05)(0) = 0$	$(0.05)(1) = 0.05$
B	$(0.20)(1) = 0.20$	$(0.20)(0) = 0$	$(0.20)(0) = 0$
C	$(0.75)(0.0625) = 0.046875$	$(0.75)(0.875) = 0.65625$	$(0.75)(0.0625) = 0.046875$

Let us now suppose that we consistently operate the following *comprehensive* rule:

- (a) whenever the score is 4, we say that the penny is of class A.
- (b) whenever the score is 0, we say that the penny is of class B.
- (c) whenever the score is 1-3, we say that the penny is of class C.

Our assertions thus constitute an exclusive and exhaustive set of events to each of which we can assign a probability of truthful statement:

$$(a) 0.05 ; (b) 0.20 ; (c) 0.65625$$

Since the set is exclusive and exhaustive, we can apply the fundamental addition rule to derive the overall probability of truthful assertion, i.e.

$$P_t = 0.05 + 0.20 + 0.65625 = 0.90625$$

$$\therefore P_f = (1 - P_t) = 0.09375$$

If we look at the balance sheet in its entirety, as in this example, we get a new slant on the misleading implications of the word *posterior*. Our posterior probabilities are of use only when we define a rule which *limits a verdict to situations in which a certain score class henceforth turns up*. Thus we pass by easy stages from thinking of the probabilities so defined as probabilities referable to one class of situations in our *total* experience to thinking of them as probabilities which are definable only in terms of what we have *already* experienced. That the last step is a step in the dark is evident when we explore the possibility of formulating rules which are not comprehensive in the sense that the foregoing rule is comprehensive.

If we confine ourselves to statements to the effect that the penny is *not* phoney (class A or B), we might operate within the framework of the rule: say that the penny is normal whenever the score is 1-3 and otherwise say nothing. We then restrict our procedure to 0.65625 of all the trials we encounter, and our balance sheet then shows that the probability of

correct statement in conformity with the understanding that we operate the rule consistently is what common sense tells us, viz. :

$$P_{t.c} = \frac{0}{0.65625} + \frac{0}{0.65625} + \frac{0.65625}{0.65625} = 1 \text{ and } P_{t.c} = 0$$

Let us now examine the consequences of another rule of restricted applicability: say that the penny is *not* normal if the score is either 0 or 4, and otherwise say nothing. The relevant figures of our complete balance sheet are then

	Score 0 or 4	Score 1-3
A	0.05000	0
B	0.20000	0
C	0.09375	0.65625
Total	0.34375	0.65625

The so-called posterior probability of the event that the penny is of class C when the score is 0 or 4 is here

$$\frac{0.09375}{0.34375} \simeq 0.273$$

Within the framework of the rule last stated, this represents the probability of *erroneously* stating that the penny is phoney, i.e.  $P_{f.ab} \simeq 0.273$ . There is, of course, no inconsistency between the statements  $P_{f.ab} \simeq 0.273$  and  $P_{f.c} = 0$ , since each endorses statements referable to a rule applicable to the outcome of only a limited class of trials, and each as such is referable to a different body of experience. The consistent operation of either rule last stated has at least a limited usefulness only because we know that each prior probability of the balance sheet is referable to an existential population at risk. If all the pennies in the bag were phoney, the first of the two would be useless, being inapplicable. The same would be true of the second if all the pennies were normal.

So far, we have assumed that we know how many pennies of each sort the bag contains. Let us next assume that we are dealing with a situation in which we know that there are

three sorts of pennies in the bag without knowing how many of each. To explore what we can then truly say about the long-run value of the 3-fold comprehensive rule of p. 153 we need to formalise it:

$$\begin{aligned} P_t &= P_a(1) + P_b(1) + P_c(0.875) \\ &= P_a + P_b + P_c - (0.125)P_c = 1 - (0.125)P_c \\ \therefore P_t &= (0.125)P_c \end{aligned}$$

Now  $P_c$  by definition lies in the range 0 — 1 inclusive, i.e.  $P_c \leq 1$

$$\therefore P_t \leq 0.125$$

If we endorse the principle of insufficient reason, we shall set  $P_a = P_b = P_c = \frac{1}{3}$ , and derive  $P_f = 0.041\bar{6}$ , a result which will be inconsistent with the entries of our balance sheet if the composition of the bag is as specified above; but the fact that we invoke the principle implies that we cannot certainly exclude such a possibility. If we speak of the probability ( $P_f$ ) of wrong assertion when we operate the rule consistently, as the *uncertainty safeguard* of the rule, we may thus say of situations to which Bayes's theorem is factually relevant:

(i) when we can assign the true values of the prior probabilities, we can frame a rule to which we can also assign an *exact* uncertainty safeguard;

(ii) when we cannot assign the true values of the prior probabilities, we can merely state an upper limit to the uncertainty safeguard of the rule which we operate.

We are now in a position to cover the situations specified by (iii) and (iv) on p. 151. We have stated a rule to which we can assign an uncertainty safeguard with a specifiable *upper limit* for *all* values of the prior probabilities including as a limiting case the possibility that the value of one is unity and that of all others is zero. This covers case (iii) which is consistent with the possibility that the value of any or all except one of the prior probabilities is zero, and case (iv) of which we know in advance that the value of one is indeed unity and that of all others is zero. We can thus resolve the antinomy inherent in our initial

statement. From the viewpoint of the practical man, there is an essential factual distinction between (i) and (iv), because (i) alone is literally an experiment carried out in two stages. From that of the mathematician, (iv) is merely a special case of (i), because the first stage of the experiment is irrelevant, if we assign unity as the value of one of the prior probabilities. The conflict no longer exists if we state our rule in a form applicable to the situation regardless of any numerical values we may assign to them.

Hitherto, we have directed our attention to the formal statement of the rule without reference to its operational content, if any; but it serves no useful end to frame a rule of procedure which specifies the risk of error unless the risk of error is acceptable, i.e. does not exceed a specified limit. Thus no one would be enthusiastic about a rule which would merely ensure 50 per cent or less correct decisions and 50 per cent or more false ones. Let us then suppose that we are content that 95 per cent of our decisions will be true in the long run, if we consistently do follow a rule stated in advance of any particular trial, i.e. our uncertainty safeguard must not exceed 0.05. How shall we accomplish our task, if we choose the 3-fold comprehensive rule under discussion? For the 4-fold trial the probability of getting 0 or 4 heads when the penny is normal is  $(0.5)^3 = 0.125$ . More generally for an  $r$ -fold trial it is  $(0.5)^{r-1}$ . For the  $r$ -fold trial we may therefore write

$$P_f = P_c(0.5)^{r-1} \leq (0.5)^{r-1}$$

Since  $(0.5)^4 > 0.05 > (0.5)^5$  we must make  $r$  at least 6 to ensure that  $P_f \leq 0.05$ . In other words, we can interpret a rule consistent with the Forward Look to accomplish the end in view, i.e. to satisfy a preassigned level of acceptability, if (and only if) the rule includes *in advance a specification of the size of the sample chosen*. At the operational level this is what the classical theory demands. It cannot prescribe a rule for the division of the stakes with an assignable risk of loss or gain, unless the prescription encompasses a specification of the number of unit trials per game. We shall see later that this limitation is the core of the current controversy concerning the credentials of the significance test.

PART II

---

*The Calculus of Error  
and the Calculus of  
Exploration*





## CHAPTER SEVEN

### THE NORMAL LAW COMES INTO THE PICTURE

WE HAVE NOW behind us the grand climacteric of the classical period. We stand on the threshold of a new phase. In tracing the origins and assumptions peculiar to a *Calculus of Errors*, one part of our task will be to separate what are: (a) purely mathematical issues within a consistent calculus; (b) logical and factual issues relevant to the use of the calculus in everyday life.

The reader who stumbles on statements to the effect that Laplace or Gosset discovered, and Liapounoff or R. A. Fisher first proved rigorously, such and such a theorem is apt to conclude that: (a) proof in this context has something to do with the validity or usefulness of the application; (b) the notion implicit in the mathematician's conception of rigour is necessarily connected with the relevance of the axioms of the calculus to the world's work. The need to draw this distinction emerges so soon as we first invoke the continuum as a useful computing device to sidestep the labour of deriving a solution which, if exact in every sense, is inconsistent with the assumption of continuity.

In the eighteenth century a major preoccupation of mathematicians was the derivation of infinite series which facilitate computations for astronomy and navigation. In that context the exploration of approximate methods for solving with as much accuracy as need be summations of terms of an unwieldy binomial series involved no departure from the outlook of the Founding Fathers. It was thus that de Moivre, Bernoulli and Laplace hit on the result sometimes pretentiously called the theorem of Laplace, viz. that the normal curve closely tallies with the contour of the histogram or so-called frequency polygon for successive terms of the binomial  $(q + p)^r$  when  $r$  is large in comparison with the reciprocal of both  $p$  and  $q = (1-p)$ .

This demonstration is a piece of pure algebra. It has no necessary relevance to the uses of a calculus of probability, if

only because its invocation as a device for quadrature of sufficient precision for practical purposes demands numerical investigation. Which of several more or less elementary procedures the pure mathematician regards as a more or less rigorous justification for the conclusion that the theorem is true in the limit is of little interest in practical situations. Our concern is then to assign particular numerical values to  $r$ ,  $p$ ,  $q$  sufficient to ensure the precision we demand of the computation. Here it will suffice to say that the correspondence over a range of 95 per cent of the area of the histogram will be as close as we are likely to want it to be for all values of  $r$ ,  $p$  and  $q$ , if  $rp$  and  $rq$  both exceed 10.

Empirical investigation alone can justify our confidence in any empirical use to which we may put the result stated. In any event, its derivation as a limiting case has its proper domain in the natural history of the binomial theorem rather than in that of the art of gambling or of naturalistic applications of statistical theory. In so far as an applied theory of probability makes use of the binomial theorem, it does so when the topic is an enumerable collection of objects, i.e. when the postulate of continuity is factually irrelevant to the situation. To Laplace and to his immediate predecessors it would scarcely seem necessary to make this concession explicit, because the modern concept of the number continuum had not taken shape. When it did take shape the calculus of probability provided a playground in which the pure mathematician could develop a calculus on the basis of initial assumptions more general than those from which the validity of the calculus in the domain of scientific research derives its sanction. By then, as we have now seen, the doctrine of inverse probability had firmly but unobtrusively installed a continuum of suppositious prior probabilities in the sepulchral regions where mental images have their place of abiding.

Part of the task of putting the new orientation of statistical theory into its rightful historical perspective is thus to disentangle axioms introduced to exercise the ingenuity of mathematicians from axioms which have some connexion with the practical utility of their pursuits. Two circumstances in particular conspired to give the infinitesimal calculus a status

# THE NORMAL LAW COMES INTO THE PICTURE

which the thought of the classical period could scarcely endorse. One I have mentioned, the introduction by Laplace of the Type II Beta Function as the keystone of the doctrine of inverse probability. I cannot do better than quote Brunt (*The Combination of Observations*) with respect to the other:

Gauss (Werke, IV, p. 116) took Bessel's reduction of 470 observations of the right ascensions of Procyon and Altair made by Bradley, and compared the distribution of errors with the theoretical curve obtained by evaluating  $h$  by the above formula. He calculated the numbers of observations whose errors should be numerically between  $0''.0$  and  $0''.1$ , between  $0''.1$  and  $0''.2$ , etc., and compared them with the actual numbers obtained from Bradley's observations. The results are given in the following table.

<i>Errors</i>	<i>Theoretical number</i>	<i>Actual number</i>
$0''.0$ to $0''.1$	94.8	94
$0''.1$ $0''.2$	88.8	88
$0''.2$ $0''.3$	78.3	78
$0''.3$ $0''.4$	64.1	58
$0''.4$ $0''.5$	49.5	51
$0''.5$ $0''.6$	35.8	36
$0''.6$ $0''.7$	24.2	26
$0''.7$ $0''.8$	15.4	14
$0''.8$ $0''.9$	9.1	10
$0''.9$ $1''.0$	5.0	7
above $1''.0$	5.0	8

The table shows a remarkable correspondence between the theory and the observational data. There is, however, a slight discrepancy in the number of large errors, the number occurring in practice exceeding the theoretical number.

The major contributions of Gauss relevant to our present theme appeared in the twenties and thirties of the nineteenth century. In the development of the theory associated with his name a contribution of outstanding interest is the *Grundzuge der Wahrscheinlichkeits Rechnung* of Hagen (1837), published two years after Quetelet's notorious *Essai de Physique Sociale*. Hagen

sought to explain the remarkable correspondence between the figures of the foregoing table and the requirements of what we now call the normal law as the limit of a binomial series of frequency terms in conformity with a situation to which the classical theory of probability is truly relevant. His explicit postulates are as follows:\*

1. An observed error ( $m$ ) is the algebraic sum of a very large number of minute elementary errors of equal magnitude ( $\epsilon$ ).

2. Positive and negative errors ( $+\epsilon$  and  $-\epsilon$ ) occur with equal frequency in the long run. From this it follows that:

(a) if  $(r - x)$  out of  $r$  elementary errors whose algebraic sum is  $m$  are positive and  $x$  are negative

$$(r - x)\epsilon + x(-\epsilon) = m = (r - 2x)\epsilon \quad . \quad . \quad (i)$$

(b) the mean value of  $m$  in an indefinitely large number of trials is zero, whence if  $v_t$  is the true value of the measurement and  $v_o$  the observed one:

$$v_t = v_o + m ; E(v_o) = E(v_t - m) \\ \therefore E(v_o) = v_t - E(m) = v_t \quad . \quad . \quad (ii)$$

If we wish to bring the above into harmony with the classical calculus of probability, we need to invoke the principle of *equipartition of associative opportunity*, i.e. that of the *Irregular Kollektiv*. We must thus add a third implicit postulate to the foregoing:

3. If the assumed fixed value of contributory errors to the observed error  $m_k$  at the  $k$ th trial is  $r$  the probability ( $P_{x,k}$ ) that  $m_k = (r - 2x)\epsilon$  is *independent* of  $k$ , being uniquely determined by  $x$ , i.e.

$$P_{x,k} = r_{(x)} \cdot 2^{-r}$$

Given (1) — (3) above Hagen's argument is translatable in terms of the gaming idiom of the founders of classical theory by reference to the following model. We toss an unbiased

\* Since we do not here initially invoke the continuum in the foregoing statement, it is pertinent to quote Brunt (p. 11):

"Hagen based his proof on the assumption that an accidental error consists of the algebraic sum of a *very large number* of infinitesimal errors of equal magnitude, and as likely to be positive as negative." (Italics inserted.)

penny  $r$  times. At each unit trial we record the score as  $+\epsilon$  if it falls head upwards and  $-\epsilon$  if it falls tail upwards. If  $x$  is the taxonomical\* tail-score, the total score ( $m$ ) in the domain of representative scoring adopted in this context is:

$$(r - x)\epsilon + x(-\epsilon) = m = (r - 2x)\epsilon$$

The mean of the  $x$ -score distribution is  $\frac{1}{2}r$  and that of the  $m$  distribution is  $E(m) = 0$ . The variances of the two distributions are respectively:

$$\sigma_x^2 = \frac{r}{4} \text{ and } \sigma_m^2 = 4\sigma_x^2\epsilon^2 \quad . \quad . \quad . \quad (iii) \dagger$$

The probability of getting a tail-score  $x$  is

$$P_x = \frac{r!}{x!(r-x)!} 2^{-r} \quad . \quad . \quad . \quad (iv)$$

If  $r$  is very large we may write:

$$P_x \simeq \frac{1}{\sigma_x \sqrt{2\pi}} \exp \frac{-(x - \frac{1}{2}r)^2}{2\sigma_x^2} \quad . \quad . \quad . \quad (v)$$

With appropriate change of scale the ordinate equation of the corresponding fitting curve of the  $m$ -distribution is

$$y_m = \frac{1}{\sigma_m \sqrt{2\pi}} \exp \frac{-m^2}{2\sigma_m^2} \quad . \quad . \quad . \quad (vi)$$

The probability of getting a value of  $m$  in the interval  $m \pm \frac{1}{2} \Delta m$  is then

$$\Delta P_m \simeq \frac{1}{\sigma_m \sqrt{2\pi}} \cdot \exp \frac{-m^2}{2\sigma_m^2} \cdot \Delta m$$

\* The distinction between taxonomical and representative scoring here and elsewhere made is the writer's own usage. Taxonomical scoring enumerates objects or classes, representative scoring records totals and averages of assigned score values.

† The scale of the  $x$ -distribution is  $\Delta x = 1$ . For the 2-fold toss we may score  $(-\epsilon, -\epsilon)$ ;  $(-\epsilon, +\epsilon)$ ,  $(+\epsilon, -\epsilon)$ ;  $(+\epsilon, +\epsilon)$  so that possible values of  $m$  are  $-2\epsilon, 0, +2\epsilon$ . Similarly,  $m = -3\epsilon, -\epsilon, +\epsilon, +3\epsilon$  for the 3-fold toss. Thus the scale of the  $m$ -distribution is  $\Delta m = 2\epsilon$ . To derive (vi) we note that the area of the histogram column of height  $y_x$  is  $y_x \Delta x = y_x$  for the interval  $x \pm \frac{1}{2} \Delta x$  on the  $x$ -score scale. On the  $m$ -score scale the width of the interval is  $\pm \epsilon$  for the corresponding score  $m = (r - 2x) \epsilon$ . If  $y_m$  is the height of the column whose area is  $y_x \Delta x = y_x$ , we may write  $y_x = y_m \Delta m$  whence  $y_x = 2\epsilon \cdot y_m$  in the derivation of (iii).

If we seek to relate the theory of the model to practice, we do not know the numerical values of  $r$  or of  $\epsilon$ . Hence we do not know the value of  $\sigma_m$ . In accordance with (iii) above, the earlier writers put:

$$\frac{1}{2\sigma_m^2} = h^2 = \frac{1}{2r\epsilon^2} \quad . \quad . \quad . \quad . \quad (vii)$$

Thus the ordinate equation assumes the form:

$$y_m \simeq \frac{h}{\sqrt{\pi}} e^{-h^2 m^2} \quad . \quad . \quad . \quad . \quad (viii)$$

In this expression, the so-called *precision index* ( $h$ ) is a constant to be determined approximately by the data of a sufficiently large experiment. The probability that an error will be as great as  $+a$  is

$$P(m \leq a) \simeq \frac{h}{\sqrt{\pi}} \int_{-\infty}^a e^{-h^2 m^2} . dm \quad . \quad . \quad . \quad (ix)$$

For purposes of tabulation we may write  $hm = t$ , so that:

$$P(m \leq a) \simeq \frac{1}{\sqrt{\pi}} \int_{-\infty}^{ha} e^{-t^2} . dt \quad . \quad . \quad . \quad (x)$$

One speaks of the table of the definite integral (x) as the table of the *Error Function*, sometimes written *Erf(ha)*, still cited in handbooks for the use of physicists, surveyors and astronomers. It is, of course, equivalent to the table of the normal integral of unit variance and of zero mean with suitable change of scale. For the latter our standard score ( $k$ ) is

$$k = \frac{a}{\sigma_m} \text{ and } P(m \leq a) \simeq \frac{1}{\sqrt{2\pi}} \int_0^k e^{-\frac{1}{2}c^2} . dc \quad . \quad (xi)$$

The definite integral on the right of (xi) is that of the normal curve of unit variance as tabulated for use in modern textbooks of statistical theory. It is more convenient, since the unbiased estimate ( $s_m$ ) of the unknown  $\sigma_m$  has a simpler relation to the empirical data. By definition  $v_0 = v_t - m$  as in the derivation of (ii) above. Thus the distribution of  $m$  and  $v_0$  the observed

value of the measurement involves only a shift of origin when the variance of the distribution of the individual measurements ( $v_0$ ) is that of the distribution of errors ( $m$ ). The mean of the former is  $v_i$  of which the unbiased estimate based on  $n$  successive observations is

$$M_0 = \frac{1}{n} \sum_{x=1}^{x=n} v_{0,x} \quad . \quad . \quad . \quad . \quad . \quad (xii)$$

The unbiased estimate of the unknown variance of the distribution is:

$$s_m^2 = \frac{1}{n-1} \sum_{x=1}^{x=n} (v_{0,x} - M_0)^2 \quad . \quad . \quad . \quad (xiii)$$

When we test the theory we rely on the hope that  $\sigma_m \simeq s_m$  if  $n$  is large in accordance with the first postulate of Hagen's so-called proof; and we may then invoke Bernoulli's theorem to justify our confidence that the error involved in the approximation will very rarely be sensible if  $n$  is sufficiently large.

Our statistical model gives a *possible* explanation of a correspondence which Bradley's data illustrate. This does not conclusively prove that the two explicit postulates are correct; and it presupposes that the approximation in substituting (v) for (iv) is valid. The numerical correspondence between the latter is indeed very close if  $r > 16$ . The derivation given above implies that there will be  $(r+1)$  different values of  $m$  whose relative frequencies tally with successive terms of  $(\frac{1}{2} + \frac{1}{2})^r$ . We shall therefore not expect good correspondence unless the recorded observation may be referable to at least 17 different scale divisions of the instrument. In the example cited from Gauss, we have before us 21.

Such a spread is by no means a universal experience and the background of the experiment Gauss cites is therefore instructive. A modern observatory is a highly complex mechanism in which large numbers of cog wheels engage in the process of setting an instrument in position for any single recorded observation. Thus all the combined effect of sources of relevant variability may be considerable when we have eliminated all systematic errors. In many types of experiment this is manifestly untrue. If we repeatedly titrate from the same

solution with the same burette and with the same pipette, we do not expect successive observations performed with competence to deviate by more than one scale division—or at most two—on either side of a central value. Thus the relevance of Hagen's postulates or of any deductions we may draw from them to an experimental situation depends on the nature of the latter, and no particular verification in one domain of experiment is necessarily relevant to any other.

Let us therefore examine the postulates more closely. Our Hagen model embodies three quite arbitrary assumptions which call for separate consideration :

- (a) the elementary errors are of *equal* magnitude ;
- (b) the particular combination of elementary errors collectively equivalent to the deviation of an observed from its true value is the outcome of a native *randomising* process ;
- (c) negative and positive errors occur with *equal frequency* in the long run.

Of the first and second of these, it suffices to say that neither is amenable to empirical verification. The second implies a symmetrical distribution of observed measurements repeated on a sufficiently large scale. The preceding tabulation does not in fact exhibit whether the original distribution was more or less skew or not.

Though together *sufficient*, no one of the two explicit postulates is indeed a *necessary* condition that the normal curve should be a good descriptive device for the distribution of repeated measurements ; but (b) and (c) are of basic importance to the classical theory of error *per se*. If we reject (a) as unnecessarily arbitrary, we can conserve the peculiar status of the arithmetic mean in the Gaussian system only by rephrasing (c) in a more general form, viz.: *the mean of the distribution of elementary errors is zero*. It then follows that their combined effect will cancel out in the mean of a sufficiently large number of observations, i.e. that the arithmetic mean of our experimental values ( $v_0$ ) approaches the unknown putatively true value ( $v$ ) more and more closely as we enlarge the number of observations recorded. This is implicit in any intelligible theory of error *sensu stricto*.



That the third postulate is irrelevant to the derivation of the so-called law itself is deducible from elementary considerations which antedate the Gaussian Theory. In our model set-up, there is equal probability of scoring heads (+  $\epsilon$ ) or tails (−  $\epsilon$ ) at each toss contributory to the  $r$ -fold trial. In the customary symbolism, we may denote by  $p$  and  $q$  respectively the probabilities of one or other event in the unit trial. We should then write more generally for (iv) and (iii) above respectively

$$P_x = \frac{r!}{x! (r-x)!} p^{r-x} q^x \text{ and } \sigma_x^2 = rpq$$

As already stated (p. 160), numerical investigation then shows that the approximation specified by (v) holds good if  $rq$  and  $rp$  both exceed 10. The rest of the demonstration is valid with due regard to the reinterpretation of  $\sigma_m$  in terms of  $\sigma_x$  defined as above.

If we regard the number of elementary errors contributory to the total error and the probabilities  $p$  and  $q$  as consistent with the criterion stated immediately above, the first and second explicit postulates of our model set-up together furnish a sufficient condition for reliance on the normal curve as a descriptive device. That the third is not a necessary one is less obvious than its arbitrariness. Advanced textbooks of statistics show how it is possible to arrive at the same result without assuming that the elementary errors are of equal magnitude. Laplace foreshadowed this derivation known as the *Central Limit Theorem* which has played a very large part in the background of what we shall later call the *Quetelet mystique*. Kendall defines the theorem as follows: "under certain conditions the sum of  $n$  independent random variables distributed in whatever form tends, when expressed in standard measure to the normal form as  $n$  tends to infinity."

The search for what mathematicians call a rigorous proof of the theorem has provoked much discussion and leads us into unnecessarily deep waters, if we arbitrarily assume a truly continuous distribution of elementary errors of different magnitude. If we stick to the firm ground of a model we can visualise, we shall suppose that a lottery wheel has some number of sectors  $N$  each labelled with some score value

$\epsilon_{0.1}$ ,  $\epsilon_{0.2}$ ,  $\epsilon_{0.3}$ , etc., not necessarily of the same sign. We shall spin the wheel  $r$  times, record at each spin the particular value  $\epsilon_{0.x}$  of the sector which stops against a fixed pointer and define the score-sum ( $s_0$ ) of the  $r$  single trials as our total "error" of observation, i.e.

$$s_0 = \sum_{x=1}^{x=r} \epsilon_{0.x} \quad . \quad . \quad . \quad . \quad . \quad (xiv)$$

Since we might record at a single trial  $r$  values of the numerically highest negative or positive value of  $\epsilon_{0.x}$ , the range of the distribution of  $\epsilon_{0.x}$  will be small in comparison with that of  $s_0$ , if  $r$  is large. *En passant* we note that we have dispensed with the second postulate, as given above, or in the alternative form which alone endows the arithmetic mean with the special meaning it enjoys in the Gaussian theory, i.e. we do not assume that the distribution of elementary errors is symmetrical about zero mean or even that the mean is zero.

The student, who has an elementary acquaintance with the notion of *moments* as descriptive measures of the contour of a distribution, will be content if we assume that two distributions are identical when all corresponding moments about the mean when expressed in standard measure are identical. It then suffices to establish the following result which relies only on elementary algebra, if we invoke our implicit assumption, i.e. statistical independence of unit trials: the standardised mean moments of the distribution of the  $r$ -fold score sum referable to any discrete score distribution approach those of the normal distribution of unit variance more and more closely as we increase the value of  $r$ .\*

The *Central Limit Theorem* occupies a pivotal, if unobtrusive, place in statistical thought from the time of Laplace to that of Liapounoff (1901) who first gave what a pure mathematician of today would concede to be a rigorous proof. This may be partly because it proceeds from less arbitrary postulates than those of Hagen to a possible explanation of why the normal law gives a tolerably satisfactory description of the outcome of repeated observations in some types of physical experimentation. If so, it illustrates how much relevance we may sacrifice

\* See Appendix I.

or greater generality in the algebraic sense of the term. Unless we assume a distribution of elementary errors about zero mean, we have no justification for interpreting the limiting value of the mean of our observations or that of any other parameter of their distribution as the *true one*; and we have no solid foothold for a concept of error, if we jettison the concept of a true value as the goal of our endeavours. Whatever bearing on scientific enquiry the theory of probability may or may not rightly have, no one fully informed will deny that: (a) the attempt to formulate a stochastic theory of error was its first considerable claim to recognition as such; (b) the founders of the theory took the view last stated, as did their immediate successors. Thus Brunt (1917) says:

This relation\* is of such importance that it is necessary to consider it in some detail. The residuals  $v_1, v_2$ , etc., are the deviations of the observed values from the A.M., and if the A.M. could be definitely regarded as the *true* value of the unknown, the (M.S.E.)<sup>2</sup> ought to be equal to  $\frac{[vv]}{n}$ . (*Italics inserted.*)

On any terms, the Gaussian theory of error, and the vast superstructure erected thereon, is indeed an entirely meaningless algebraic exercise, unless we conceive the attempt to arrive at a *true* value as the reason for making repeated observations. What endows the central limit theorem with a more special interest is that it tempts us to carry into quite a different domain from that of the combination of observations the formal algebra of the Gaussian theory without probing into its relevance too deeply. This development begins with Quetelet. It was he who first drew attention to the distribution of heights of adults in a comparatively homogeneous population, and advanced the proposition that its form is a manifestation of (his words) the *binomial law*. In short, nature is an urn which shakes out numbered *billets* at random. The individual is a packet of such *billets*. His or her score—height or whatnot—is the numerical sum of the numbers of the constituent *billets*.

We may seek a rationale for Quetelet's so-called law of

\* Unbiased estimate of variance. In the symbolism of Brunt [vv] at the end of the citation stands for the sum of the squares of the residuals  $v_1, v_2$ , etc.

height at very different levels. If we disregard the role of environment, we can nowadays invoke stochastic principles from the domain of the genetical theory of populations. With some plausibility we can then identify a pool of elementary entities with a system of supposedly *additive* multiple factors each assigned a hypothetical score  $\epsilon_{0,x}$  in an assumed framework of random mating. To give this hypothetical score any substance from a genetical viewpoint, we have to think of it in terms of gene substitution, i.e. against the background of a reconstructed ancestral "wild type." In a standard environment, we thus conceptualise each elementary score as a deviation of so many units of relevant measurement. However, the ancestral wild type which we invoke is beyond recall.

Quetelet's following, in particular Galton and K. Pearson, had no such empirically validated stochastic theory to discipline their speculations. They took over from Darwin a vague particulate hypothesis, which leaves us with no origin of reference for the hypothetical elementary score value  $\epsilon_{0,x}$  unless we arbitrarily define it as the *norm* of the population. In that event, we become involved in the comic obligation to decide arbitrarily what parameter of the population distribution is the true norm. Indeed, discussion about whether to plump for the arithmetic mean, the geometric mean, the harmonic mean, the mode or the median filled pages of statistical textbooks issued in the heyday of the singular cult of biometrics. It should be—but alas is not even yet—needless to say that the issue is meaningless. Only the assumed existence of a true value which we may hope to approach more closely by taking the mean of an ever larger number of observations endows the arithmetic mean or any other parameter of their distribution with a unique semantic content in contradistinction to what claims it may have if we wish to specify a descriptive curve in the most convenient way.

That it is still necessary to state this truism is evident from the fact that contemporary literature on significance tests follows the practice of Yule by ascribing a special status to the probability that the deviation of an observed measurement from the true mean of the parent population [*sic*] will attain a certain magnitude. To keep up the illusion we still speak of

the mean as the *expected* value of a statistic and denote by  $E$  (or  $\epsilon$ ) the operation of extracting the arithmetic mean of a sampling procedure; and we may readily forgive the pure mathematician who walks into the trap, when the same symbolism serves to specify both a *deviation* from a non-existent norm and an *error* in determining the *true value* of a measurement. What is not easy to understand is how this formal identity can hoodwink anyone accustomed to naturalistic pursuits.

This will be clear if we examine what we really mean when we say that 75 per cent is the *expected* proportion of peas with yellow seeds from *a* mating of yellow-green hybrid parent plants. The italicised indefinite article in the last sentence is operative. The calculus of probability does not and cannot entitle us to speak of *a* mating. In this context it can merely provide the basis for a verifiable theory about a particular class of matings *in the long run*. It does not even justify the statement that we shall encounter 75 per cent more often than any other proportion of yellow peas in an indefinitely protracted number of trials. Unless the sample size ( $r$ ) in the sequence of trials is an exact multiple of 4 an observed proportion of 75 per cent yellow seeds is an event we can never expect to encounter in *any* such trial.

The last consideration puts the spotlight on what is perhaps the most puzzling consequence of the formal identification of a natural deviation from a suppositious norm with an error as Gauss conceived it. In a scrupulously careful repetition of matings from inbred stocks no error as Gauss conceived error need arise, if we employ reliable mechanical aids to *enumeration*. In the system of Gauss, the deviations from the limiting value of the mean signifies man's fallibility in the apprehension of natural law, and as such we deem it to be irrelevant to the precise formulation of a law of nature. In the so-called law of height our observed deviations from the norm, if recorded with sufficient precision, embody the content of the so-called law itself.

In the last two paragraphs I have deliberately anticipated an intrusion of the calculus of probability into the domain of experimental science at a different level from that of Gauss,

and after a lapse of nearly a century. I have felt it necessary to do so to get into focus an issue raised by Brunt (*vide infra*). In one sense, the Central Limit Theorem offers us the algebraic apparatus for an explanation of a class of occurrences, but not as men of science commonly use the word *explanation*. It is evidently consistent with any particulate theory of inheritance, right or wrong, in situations which admit of explanation in terms of inheritance alone. It might well be likewise possible with a little ingenuity to bring earlier speculations which invoke humours, fluids, essences *et hoc genus omne* within the scope of its catholic terms of reference; but what have we gained by doing so? Can it specifically lead us to a deeper understanding of nature, by disclosing hitherto unsuspected phenomena? The answer is that any confidence in stochastic reasoning we now derive from the advancement of knowledge concerning heredity arises from circumstances to which the Gaussian theory of error is irrelevant.

The influence of Quetelet on subsequent thought is the more remarkable because the legacy of absurdities he bequeathed to posterity were fully recognised by writers on probability in the middle of the nineteenth century; and physicists who endorse the Gaussian theory of error have been highly critical of the uses to which Quetelet's followers have put it. In the preface of his book already cited, Brunt expressly warns his readers against the folly of confusing unavoidable human error with natural variation; but the cardinal absurdity of doing so is condoned by implication in any current treatise on statistical theory for research workers in the biological and social sciences.

It is therefore both gratifying and instructive to recall the engaging summary of Quetelet's views by Bertrand\* in his *Calcul des Probabilités* (1888). Bertrand does not fail to anticipate how heavily the superstructure of regression theory erected on the foundations Quetelet laid leans on a scaffolding of Platonism:

... The world of universals seemed talked out and quite forgotten. M. Quetelet, without reviving this ancient problem, seriously believes he has resolved it and, in a book stuffed full of judiciously

\* I must thank Dr. Richard Padley for a translation which does justice to the Voltairean flavour of the original.

collected facts, would have us accept a precise definition of the word Man, independently of human beings whose particularity can be considered accidental. With as little discussion as subtlety, the painstaking author defines his specimen, attributing to him the arithmetic mean of every element that varies from one man to another. After a survey, for example, of the heights of 20,000 soldiers, the mean has been determined at 1 m 75; such then is the height of the average man. Around it in the scale of measurement are grouped greater or lesser statures, exactly graduated according to the law of error. Nothing distinguishes the heights of the conscripts from 20,000 successive measurements which an incompetent observer would have taken on a man 1 m 75 tall if we suppose the work to have been carried out with instruments which, though crude enough, were corrected for any constant error.

In this comparison M. Quetelet sees an identity. Our inequalities of height are, in his eyes, the result of inept measurements taken by Nature on an immutable model in whom alone she reveals her secrets. 1 m 75 is the normal height. A little more makes no less a man, but the surplus or deficit in each individual is nature's error, and thus monstrous.

Abelard, if set to this disputation, would have presented the argument in formal terms, but such subtlety no longer holds sway. Wandering over the schoolmen's ancient battlefield, M. Quetelet has fallen in with neither ally nor foe.

The Thesis has, however, more than one inconvenience. The ideal man, we shall say, represents in all things the arithmetic mean of humanity. This sounds very simple and very clear, but how are we to compute these measurements defined within the limitations of ruler and compasses. Mean height of head, for example, can be computed in two ways; we can take the mean of head lengths, or for each individual, the relationship between head and body length and then the mean of these ratios. The results are different; how should they be brought together.

The difficulty is serious and shipwreck certain. To show this in a model situation let us examine the mean of two spheres. The first has unit radius and we shall choose a scale such that surface area and volume are also unity. The second, I will suppose, has radius 3 and, necessarily, will have a surface area of 9 and volume 27. Means of 2, 5 and 14 are incompatible; a sphere of radius 2 would have a surface area 4 and volume 8 exactly; no concession is possible, a sphere can have no other shape. Men's shapes unfortunately can vary, and M. Quetelet profits therefrom. By combining the mean weight of 20,000 conscripts with their mean height, we should

produce an absurdly fat man and, whatever Reynolds might have said, a poor model for an artist. This eminent painter, in his Lectures on the Fine Arts preceded Quetelet in setting up the average man as the type of perfect beauty. If such were the case, suggested Sir John Herschel, ugliness would be the exception. I cannot follow his argument. The individual traits of perfect beauty would not be rare; indecorously jumbled together they would lack merit. Grace stems from harmony. Chance doubtless would summon few elect and, despite Sir John Herschel, in the ill assorted assembly, if ugliness formed the exception, the grotesque would become the rule.

In the body of the average man our Belgian author sets an average soul. To summarize the moral qualities it is necessary to cast 20,000 characters in one. The specimen man would be without passion or vice, not foolish, not wise, not ignorant, not knowledgeable, generally dozing, for it is the average between wakefulness and sleep; he would answer neither yes nor no, and be in all things mediocre. After remaining alive for 38 years on the average ration of a healthy soldier, he would die, not of old age, but of an average disease which statistics would discover for him.

But for Galton, this and other epitaphs on the views of Quetelet by exponents of the theory of probability in the latter half of the nineteenth century should have buried the author of the *Essai* with appropriate honours. *Natural Inheritance* appeared in 1889, a year after Bertrand's *Calcul*. Among other diversions of a country gentleman Galton took to photography, and was able to give verisimilitude to the Quetelet norm by producing composite snapshots of the judge, the criminal and the leading counsel as visual aids to the hitherto nebulous domain of Plato's universals. The following passage (*italics inserted*) illustrates how the Gaussian law of *instrumental error* henceforth becomes the *normal law of nature*:

I need hardly remind the reader that the Law of Error upon which these Normal Values are based, was excogitated for the use of astronomers and others who are concerned with extreme accuracy of measurement, and without the slightest idea until the time of Quetelet that they might be applicable to human measures. But Errors, Differences, Deviations, Divergences, Dispersions, and individual Variations, all spring from the same kind of causes. Objects that bear the same name, or can be described by the same



phrase, are thereby acknowledged to have common points of resemblance, and to rank as members of the same species, class, or whatever else we may please to call the group. On the other hand, every object has Differences peculiar to itself, by which it is distinguished from others.

This general statement is applicable to thousands of instances. *The Law of Error finds a footing wherever the individual peculiarities are wholly due to the combined influence of a multitude of "accidents," in the sense in which that word has already been defined. All persons conversant with statistics are aware that this supposition brings Variability within the grasp of the laws of Chance, with the result that the relative frequency of Deviations of different amounts admits of being calculated, when those amounts are measured in terms of any self-contained unit of variability, such as our Q. (pp. 54-5).*

Starting with this misconception of the proper terms of reference of a calculus of error, Galton speaks (pp. 16-17) in the following passage concerning the *Variety of Petty Influences*.

The incalculable number of petty accidents that concur to produce variability among brothers, make it impossible to predict the exact qualities of any individual from hereditary data. But we may predict average results with great certainty, as will be seen further on, and we can also obtain precise information concerning the penumbra of uncertainty that attaches itself to single predications.

Thus we learn at the end (p. 193) :

A brief account of the chief hereditary processes occupies the first part of the book. It was inserted principally in order to show that a reasonable *a priori* probability existed, of the law of Frequency of Error being found to apply to them. It was not necessary for that purpose to embarrass ourselves with any details of theories of heredity beyond the fact that descent either was particulate or acted as if it were so. I need hardly say that the idea, though not the phrase of particulate inheritance, is borrowed from Darwin's provisional theory of Pangenesis, but there is no need in the present enquiry to borrow more from it.

For the reader who does not probe too deeply into "what all persons conversant with statistics are aware" of, the outcome of the enquiry is not less arresting because the author expresses his more dramatic conclusions with becoming modesty. Elsewhere, we learn (p. 48) :

My data were very lax, but this method of treatment got all the good out of them that they possessed. In the present case, it appears that towards the foremost of the successful men within fifteen years of taking their degrees, stood the three Professors of Anatomy at Oxford, Cambridge, and Edinburgh; that towards the bottom of the failures, lay two men who committed suicide under circumstances of great disgrace, and lowest of all Palmer, the Rugeley murderer, who was hanged.

We need scarcely recall that K. Pearson's flair for ancestor worship had ample scope for self-expression in his partnership with the founder of the Eugenic cult. He took over from Galton the term *regression* and sponsored a novel extension of the Gaussian technique of statistical estimation in publications which successfully concealed the biological postulates and implications of Galton's so-called *Law of Ancestral Inheritance* behind a forbidding façade of symbolic inconsequence. An issue which was essentially mathematical thus became the battleground of a biological controversy conducted with an output of heat totally disproportionate to the illumination conferred. When the shouting and the tumult died, the older generation of mathematicians had lost interest in the origins of the concept and were all too ready to accept at face value a now allegedly indispensable "tool" of biological enquiry as a means of providing full employment for their younger colleagues. How this came about we shall see more fully, when we have looked a little more closely into the Gaussian *Method of Least Squares* and the concept of covariance which emerges therefrom.

In this context our main concern is to trace to its origins the peculiar status which the normal distribution enjoys in contemporary statistical theory. The rationale of Quetelet's binomial law, espoused by Galton and thereafter endorsed as the kingpin of the theory of regression, is the suppositious relevance of Hagen's model; and the best one can say of Hagen's model in this context is that it is an attractive expository device. Its relevance to the contour of graphs defining certain empirical distributions of measurements made on populations of living beings and its relevance to observed distributions of instrumental errors are alike amenable neither to proof nor to disproof. Thus Brunt (p. 17 *op. cit.*) concludes his exposition of

attempts of Hagen and others to provide a *proof* of the Normal Law of Error in the following salutary terms (*italics inserted*):

The proofs of the Normal Error Law given above are based on certain definite hypotheses concerning the nature of accidental errors. It has been shown that, if the accidental errors to which a series of observations is liable satisfy these hypotheses, the errors of observation will be distributed according to the normal law. The final justification of the use of Gauss's Error Curve rests upon the fact that it works well in practice, and yields curves which in very many cases agree very closely with the observed frequency curves. The normal law is to be regarded as *proved by experiment*, and *explained* by Hagen's hypothesis. When the curve of frequency of the actual errors is not of the form of the normal curve, we may safely conclude that the nature of the accidental errors concerned is not in accordance with Hagen's hypothesis.

As we have seen, Hagen's model is one of many models entirely consistent with the classical theory of risks in games of chance. Each can lay claim with as much and as little plausibility to furnish an *explanation* of the distribution of errors, when the distribution accords closely with the normal curve; but close correspondence between distributions of error in the Gaussian sense of the term is one which we shall expect to encounter, and shall indeed encounter, only in a limited class of experimental enquiries. Aside from the arbitrariness of the assumptions embodied in the specification of the Gaussian law of error, this reflection alone suffices to dispose of the misconception that the classical theory of risks can indeed furnish any rationale for error distributions in general.

When we transgress the boundaries of the domain of experimental error to contemplate the baffling intricacies of natural variation, no person conversant with the content of the classical theory of risks can condone the identification of the variety of petty influences with "accidents" which bring "variability within the grasp of the laws of chance" so defined. The belief that the *Central Limit Theorem* endorses such an identification is groundless for two reasons, each sufficient to dispose of the claim. One is the formal assumption that we are dealing with elementary scores whose gross effects are both

*independent* and *additive*, both suppositions certainly false in many situations, and especially in the Galtonian context of the nature-nurture issue. The other is that its acceptance as a law of nature excludes the possibility that natural variation interpretable in terms of an infinitude of such additive petty influences can ever be skew.

If unfamiliar with the history of statistical theory, the reader might well suppose that the foregoing citations from Galton exaggerate the role of the normal man in the superstructure of contemporary theory raised on the foundations Quetelet laid. If so, two citations should dispel them. The first is from a polemic note in *Biometrika* (Vol. 8, p. 249, 1912). Replying with characteristic vigour to a criticism to the effect that "we can only speak of a typical individual when we are dealing with one measurable feature at a time," Pearson asks:

Can Professor Lloyd have the least conception of what are the leading features of a multiple frequency surface? Has he never heard of the "mean man" of Quetelet, or of Edgeworth's defence of that "*mean man's*" *actuality*? (*Italics inserted.*)

What meaning we may attach to actuality in Pearson's Platonic universe of perceptions will appear from Edgeworth's own words in a publication entitled *Statistical Correlation between Natural Phenomena* (*Journ. Stat. Soc.*, Vol. 56, 1893). He therein promulgates the bivariate normal distribution:

Let it be required to construct a budget . . . representing the expenditure of a typical workman's family upon several articles of food, rent, etc. . . .

Here, as elsewhere, sociology may derive instruction from the experience of her elder sister, physical science. The case before us is analogous to that which Quetelet treated when he sought to construct a *Mean Man* by measuring the limbs or organs of a great number of men, and taking the mean of the measurements relating to each part as the type of that part. It has been objected to this method that the parts thus determined might not fit each other. . . .

It is with great diffidence that I venture to differ slightly from such high authorities; by submitting that their objection, though valid in the abstract, is much weakened by a circumstance which prevails *in rerum natura*, the fulfilment of the law of error. I need not

remind students of statistics that very generally the members of a species, e.g. men or shrimps, range with respect to any measurable attribute, such as the length of an organ, under a curve of which the equation is of the form  $y = Ke^{-ax^2}$ , . . . I have now to introduce a more general law of error, expressing the frequency of the concurrence between two, or more, attributes. . . . If, as in a former paper, we compare the curve of error to the outline of a *gendarme's* hat, we may now compare the surface of error to the top of a "pot" or "billicock" hat.

It is wonderful how accurately this double law of error is fulfilled in the case of animal organisms, as shown by the observations of Mr. Galton on men (Royal Society, 1888), and those of Professor Weldon on shrimps (*Ib.*, 1892); . . .

. . . There exists a mathematical, as well as an artistic, proportion between the parts of the human frame. . . .  $P_1, P_2$  being any points on the axis of  $x$ , if planes be drawn through them perpendicular to that axis, the highest points of the curves traced out on these planes by their intersection with the error-surface will lie on a plane perpendicular to the plane of  $xy$ , and passing through a certain straight line OR.

A case of this proposition, which particularly concerns us, is when  $x = 0$ . In that case the average, which is also the greatest ordinate or centre of greatest frequency, for one attribute corresponds to or, is in the long run most frequently associated with, the average, or greatest ordinate value, of the other attribute. Our hat has one rounded summit; it is not puckered up into irregular projections like the soft felt hats now sometimes worn.

Here is the answer to the Cournot-Westergaard objection that the average value of one organ may be inapt to coexist with the average value of the other organ. The exact contrary proves to be true. Considering the average of one organ, we see that the value of the other organ which most frequently in experience—most probably in expectation—is associated with the average of that one is the average of the other.

These propositions may be transferred from animal to social organisms; in virtue of the presumption that the compound law of error prevails in the latter, as well as in the former, department. This presumption is based upon these two premises: (1) The compound, as well as the simple, law of error is apt to be fulfilled by phenomena which depend upon a variety of independent elements or agencies. . . . (2) Social phenomena are largely of this character; as is shown, ( $\alpha$ ) generally by the constancy of statistics, a constancy which seems best explained by the play of an immense number of

influences whose fluctuations compensate each other; ( $\beta$ ) in particular by the prevalence of the simple law of error in social phenomena . . . which can hardly be accounted for otherwise than by such a combination of agencies as would equally tend to fulfil the compound law; ( $\gamma$ ) by actual verification in the particular case of correlation between the marks in Greek and Latin at the India Civil Service Examinations for 1874 and 1875—Candidates who are above or below the average mark in Greek prove to be above or below the Latin average to about the extent which the theory predicts.

No doubt in acting upon this presumption—as generally in applying mathematical ideas to social phenomena—regard must be had to the degree of irregularity which may be expected in the subject-matter. One abnormality which often characterises a group of quantities which cannot sink below zero, but may rise ever so high, is an elongation of the upper limits of the theoretically symmetrical curve of error. I have noticed this incident in the fluctuation of prices. . . .

We may extend the province of the foregoing argument to accommodate as many measurable or enumerable attributes as anatomists, physiologists and psychologists can specify. Our bivariate universe conceived in 3-dimensional space as a policeman's helmet then becomes a multivariate universe in the abstract domain of multidimensional geometry. It thus turns out that the *actuality* of the normal man is the distance of a point from an arbitrary origin in a non-visualisable hyperspace having an infinitude of dimensions. In the gospel of Pearson the just man made perfect is in short an arbitrarily selected constant definitive of what is actually an arbitrarily selected population.

Having identified the angelic choir in the Platonic empyrean of universals with an infinite population of the Normal Man, we must needs furnish it for his occupation, if we are to do justice to the nature-nurture issue conceived in such terms. By implication\* in his attempt to evaluate the roles of nature and nurture interpreted gratuitously, and in manifest contrariety to facts sufficiently familiar to the biologist, as additive and independent components of growth, R. A. Fisher does indeed invoke the concept of an environmental norm—the mean of

\* And explicitly in a private communication (1933) to the writer.

all environments. Within what geographical limits we shall locate the source of our data and within what geographical limits such data are available are questions to which no answer is conceivably attainable, the more so because we should need to be able to specify every variation of nurture relevant to development before we could list what data we require. Assuredly, we need not traverse the argument of Bertrand's exposure of the Normal Man to convince ourselves that the normal environment is a figment of the imagination devoid of any intelligible interpretation in the domain of conduct.

## CHAPTER EIGHT

### THE METHOD OF LEAST SQUARES AND THE CONCEPT OF POINT ESTIMATION

IN THE CLASSICAL THEORY of the division of the stakes our concern is with systems on which we can impose randomness, or at least aspire to do so, by prescribing appropriate precautions, e.g. thorough shuffling between successive deals. As we approach the climax of the classical period, there emerges with the doctrine of inverse probability an innovation which has exerted a more lasting influence and one which has provoked far less controversy. The concept of the infinite hypothetical population embraces the postulate that blindfold selection therefrom is necessarily randomwise; and this postulate is the keystone of a calculus of exploration propounded by Quetelet in the social milieu both of a new public concern for the collection of reliable statistics of human welfare and of a new impetus to formulate an acceptable regimen for the assessment of instrumental errors of observation.

That none of Quetelet's contemporary critics cast doubt on it is not remarkable in the context of the emergence of an ostensibly stochastic rationale for the combination of observations on physical phenomena. The same assumption is inherent in the latter, and enjoys the sanction of Gauss. To do justice to the Gaussian theory in its own domain, it is indeed necessary to distinguish two levels at which it invokes the calculus of probability. First, it invites our assent to the proposition that uncontrollable errors of observation are the outcome of the allocation of intangible additive components by a randomising process inherent in the process of measurement. It likewise invites our assent to the proposition that uncontrollable errors referable to consecutive gross measurements of the same entity are themselves the outcome of a self-randomising process. The first of these two propositions has been the topic of Chapter Seven. Only the second, which will be the theme of this, is strictly essential to the claims of the stochastic calculus of error. We



can get the distinction here stated into focus more clearly, if we now formulate two model situations.

Let us first imagine an urn A containing an infinitude of tickets each with a number, either negative or positive. We conceive that an umpire draws  $k$  tickets and records as the *gross trial score* the result of adding the sum of all the numbers thereon to a fixed constant ( $M$ ). At this stage the ballot is secret. The onlookers have no opportunity to scrutinise the draw. They cannot see the face of the umpire. They neither know how many ( $k$ ) tickets he draws at each trial nor what numbers the individual tickets bear. All the onlooker can observe is the outcome of a very large number of trials, viz. the frequency with which different values of the  $k$ -fold gross trial score occur. We may speak of this as the *unit sample distribution of observable errors*. In so far as it confers any special status on the mean of *all* trial scores, we must at least assume that it is symmetrical about  $M$ , i.e. that the sum of all the negative ticket numbers in the urn is numerically equal to that of all positive ticket numbers. Such is the stochastic model which endorses the first of the two propositions on which the calculus of error relies, if we also assume that the fictitious umpire shakes the urn thoroughly before each trial.

We now conceive the construction of a second urn B containing an infinitude of counters each with a number corresponding to a trial score of a  $k$ -fold draw from urn A. We shall also suppose that the relative frequencies of the counter score values tally with the unit sample distribution of gross trial scores referable to the outcome of the previous ballot. A second umpire now draws  $r$  counters from urn B. This time, we recognise the face of the umpire as that of an existent investigator. We can see how many counters he withdraws at each  $r$ -fold trial, and we can recognise the score on each of the  $r$  counters at each trial. What we cannot see is whether he shakes the urn thoroughly between trials. Such is the model the calculus of error invokes to justify the second proposition stated above, if we charitably assume thorough shaking.

Now the only reason for believing that the fictitious umpire withdrawing the *billets* bearing numbers definitive of suppositious elementary components of error does indeed shake

urn A thoroughly before each  $k$ -fold trial is that the unit sample distribution of observable errors as here defined, sometimes accords with a type of sampling distribution, the normal, deducible from assumptions consistent with the postulate of randomisation. The most we can therefore claim for our first proposition is that it is not inconsistent with anything we know about how uncontrollable—so-called *accidental*—errors arise. Nothing we know from direct observation or from background information conclusively endorses the postulate.

Whether the unit sample distribution of observable and uncontrollable errors does indeed conform to the normal prescription is irrelevant to the truth or falsehood of our second proposition; but if we invoke the classical theory to interpret the result of successive  $r$ -fold trials in sampling from urn B we must assume\* that any particular value the counter score may have at a single trial will turn up with the same frequency in juxtaposition to each possible value its predecessor or successor may have. This may well be so; but one seeks in vain for evidence of large-scale trials which confer any plausibility on the truth of the assertion. If we assume it to be true, the central limit theorem will justify our confidence that the mean distribution of the  $r$ -fold sample will be approximately normal for sufficiently large values of  $r$ ; but the mere fact that the normal curve proves to give a good fit does not suffice to justify the postulate that the process of successive measurement is random in the classical sense.

At this point the reader may rightly say that the validity of the postulate is irrelevant. What matters is whether the system behaves as if the postulate does hold good; and this contention is surely admissible, if we also accept the obligation to examine each system on its own merits. So far as the writer can discover, there has been little intensive study of the natural history of error with that end in view; but recent work of Wootton and King (*Lancet*, March 7, 1953) casts grave doubts on the propriety of using the normal curve as a descriptive device for distributions of sample means of chemical tests for human

\* The condition here stated does not suffice to prove that the endless series of the event is strictly without pattern; but its verification in large-scale trials would go far to reinforce our faith in the postulate of randomness.

blood constituents. We therefore arrive at the following conclusions: (i) the properties of classical models can give us no conclusive assurance that stochastic algebra will correctly describe the distribution of sample means; (ii) the use of stochastic algebra to assign frequency limits to errors of observation entails the obligation to undertake an *ad hoc* enquiry into the error distribution referable to any method of determination, when we do indeed invoke it.

The use of a statistical model such as urn A to interpret the approximately normal distribution of observations in certain types of physical enquiry is admittedly not inconsistent with the terms of reference of the classical theory, though the facts of the case furnish no reason for preferring one model to another; but if we ask why the normal distribution occupied so pre-eminent a place in the theory of error which Gauss and Hagen propounded, we come face to face with an innovation for which the classical theory furnishes no self-evident rationale. Though we shall later see that Gauss justified it by recourse to other considerations, the prominence it assumed in the exposition of the theory by the successors of Gauss was undoubtedly due to their belief that it endorses a new principle for combining observations in accordance with the outcome of random-wise sampling from urn B.

The new principle known as the *Method of Least Squares* signals the intrusion of a concept which we now speak of as *point estimation*, and the enlistment of the theory of probability in a domain of application different from that of interpreting error distributions of the type exhibited on p. 161 in terms consistent with stochastic models. The formulation of the method itself antedates Gauss, and many expositions of the basic assumptions current from that of Gauss to our own time interpret them *en rapport* with the viewpoint of Legendre and of Laplace. It took shape during a period of outstanding refinements in the design of measuring instruments, a circumstance which throws light on its ready welcome. An earlier generation might have lazily and cheerfully hoped that instruments of greater precision would eliminate discrepancies between successive determinations of the same entity; but the introduction of vast improvements of observatory equipment

leading to the detection of the annual parallax of a star, alone sufficed to dispel any grounds for believing that mechanical procedures could extricate the observer from the need to formulate some guiding principle or universal convention for minimising the error entailed in any method of arriving at an average. Better measuring devices reduce the range of valuation but on a more refined scale. Thus they do not necessarily reduce the number of scale divisions consistent with competent observation of the same physical dimension.

Most of us will be well content to give equal weight to every observation of a single entity, as when we adopt the arithmetic mean for successive measurements of the same physical dimensions, e.g. a height or a weight; but few, if any, of us could confidently furnish a wholly convincing justification for this preference. In many situations, few of us would indeed pause to challenge the propriety of the custom; but the issue assumes a more provocative aspect for two reasons when we combine different measurements to *estimate*, i.e. determine *as best we can*, such a single entity. For example, we may use the angle of elevation ( $a_1$  and  $a_2$ ) at each extremity of a base line of length ( $b$ ) to determine the height ( $h$ ) of a mountain. Each such combined observation entails errors of three sorts, or at least of two, if we reject the possibility that one of the three measurements ( $b$ ) is liable to error. In that event any single pair of observations will yield one and the same estimate of  $h$ , but the attempt to accomplish greater precision by making more observations will lead us to different estimates referable to different paired values of  $a_1$  and  $a_2$ . Nor do our difficulties end here. If we record  $a_1$  and  $a_2$  alternatively we come face to face with the question: shall we take the mean of each estimate based on a particular pair of values  $a_{1,r}$  and  $a_{2,r}$ , or shall we make an estimate based on the mean value of  $a_{1,r}$  and  $a_{2,r}$ ? Since the results will not necessarily tally, an intuitive and unreflective preference for the mean leaves unsolved the problem of combining observations in the best way.

This simple illustration of the problem subsumed in the title of more than one standard text on the *Combination of Observations* is pertinent to the historic context in which the theory emerges. In the half-century from 1780–1830, the introduction

of achromatic lenses, a new technique of glass polishing and new machine tools for wheel-cutting in the wake of emergent steam power contributed alike to improved design of the telescope, the microscope and the theodolite. Incident to these improvements, and coincident with the demonstration by Gauss of the normal law of error as a remarkably good fitting curve for Bradley's observations on Altair and Procyon (p. 161) came the final vindication of Kepler, the detection of the annual parallax of the star  $\delta$  Cygni by Bessel (1837-40). Hitherto the only co-ordinates of the map of the celestial sphere beyond the solar system had been angular. Star map-making now enters on a new phase in which the determination of interstellar distances calls for ever-increasing refinement of measurement.

In the same milieu, earth map-making embraces a new programme by taking advantage of new means, a new motive and a new opportunity to explore the earth's crust. Advances in the technology of the observatory immediately followed extensive surveying for the canal system and synchronised both with still more extensive surveying for the new railroads and with a concomitant efflorescence of geological enquiry. A theory of error which could lead to agreement about the best method of combining observations was thus a keenly felt need in the domain of geodesics. Indeed, textbooks of surveying give us the best insight into the use of the new principle on its original terms of reference. Thus it is not without interest that the author of the *Essai de Physique Sociale* was both an astronomer in charge of the national observatory and professor of geodesics in the Brussels military academy.

When we do combine different observations ( $x_1, x_2, x_3$ , etc.) to determine one or more physical dimensions or constants ( $p_1, p_2$ , etc.), different methods of doing so may commend themselves to common sense; and these may lead us to assign somewhat different values to what we may here call the required parameters ( $p_1$ , etc.). In the Gaussian theory of error different values of the parameters will entail different deviations of  $x_1, x_2$ , etc., from their putative true values ( $t_1, t_2$ , etc.). The probability assigned by the classical theory to the compound event of recording the particular set deemed to be a sample

chosen randomwise from all possible observations of the same sort within an assumed fixed framework of repetition will thus depend on whether we adopt one or other set of parameter values. For reasons which are by no means obvious nor exempt from criticism, Legendre and Laplace invoked this consideration to justify a method of combining observations individually liable to instrumental error in the widest sense of the term. In the thought of the period which antedates Gauss, the fundamental postulates of the method are two:

(i) of all sets of values we might choose as estimates of the parameters, we shall deem that set as best, if it assigns the highest probability to the compound occurrence;

(ii) the values which satisfy the criterion of preference so prescribed are those which make the sum of the squares of the residuals, i.e.  $(x_1 - t_1)^2$ ,  $(x_2 - t_2)^2$ , etc., as small as possible.

As shown below, (ii) is an algebraic tautology, if a normal distribution of independent errors holds good, i.e.: (a) the normal law prescribes the probability of a given deviation  $\epsilon_r = (x_r - t_r)$  of an individual observation from its true value; (b) the individual observations are themselves statistically independent. This indeed is what most early expositions of the Method of Least Squares proffer as a proof of it on the explicit assumption that the criterion of preference defined by (i) is also admissible; but the prescription of the criterion of preference embodied in (i), hereinafter referred to as the Legendre-Laplace Axiom, has no obvious connexion with the classical theory. Nor did Gauss himself invoke it. It calls for comment, because it has left a lasting imprint on the discussion of the place of *point estimation* in the applications of the theory of probability.

The derivation of the normal curve as an approximate description in accordance with the Hagen model or with any more generalised model of observed measurements which present themselves as *discrete* quantities in terms of scale divisions leads us to an expression for the probability of a

deviation ( $X$ ) of a particular observation from its mean value on unit scale as

$$P_x = \frac{h}{\sqrt{\pi}} e^{-h^2 X^2}$$

For a set of independent deviations ( $U, V, W$ ) subject to the *same dispersion* due to the same sources of so-called accidental error, we may regard  $h$  as constant and the probability of the particular sequence of values as:

$$P_s = K \exp - h^2 (U^2 + V^2 + W^2)$$

This expression will be a maximum, if the sum of the square deviations ( $U^2 + V^2 + W^2$ ) is a minimum. We may write the sum more explicitly in terms of the actual observations  $u, v, w$  and their corresponding true (long-run *mean*) values as

$$(u - M_u)^2 + (v - M_v)^2 + (w - M_w)^2 = E_s$$

The values of  $M_u, M_v, M_w$  which will confer on  $P_s$  its maximum value must thus satisfy the identities

$$\frac{\partial E_s}{\partial M_u} = \frac{\partial E_s}{\partial M_v} = \frac{\partial E_s}{\partial M_w} = 0$$

If all the observations refer to the same quantity, we may put  $M = M_u = M_v = M_w$  and

$$\frac{\partial E_s}{\partial M} = 0$$

$$(2M - 2u) + (2M - 2v) + (2M - 2w) = 0$$

$$\therefore M = \frac{u + v + w}{3}$$

This result merely states that the mean is the *best* value of independent observations on the same quantity, if the method of least squares is the method which leads us to the best value. It throws no light on what we mean by *best* in this context.

THE METHOD OF LEAST SQUARES. That the mean is an *un-*

*biased estimator*\* of the true value when each observation is referable to one and the same physical dimension is deducible from elementary considerations without invoking the normal or any other distribution of errors. This we shall presently see; but some readers may first wish to familiarise themselves with the type of physical situation in which the physicist invokes the method of least squares itself.

The sort of situation which first commended the use of the method of least squares to the contemporaries and immediate successors of Gauss arises when we seek to combine information from observations referable to *two* different quantities. A simple illustration from elementary physics will suffice to indicate the problem of the combination of observations at this level. We shall suppose that we wish to determine the resistance ( $x$ ) of a bridge wire, in which event our external circuit will make a fixed contribution ( $y$ ) to the measured resistance ( $m$ ). If we could determine the latter without error it would suffice to make two determinations, one ( $m_1$ ) referable to a particular fraction ( $f_1$ ) of the length of the bridge wire included in the circuit and a second ( $m_2$ ) referable to a different fraction ( $f_2$ ). We should then have

$$m_1 = f_1x + y ; m_2 = f_2x + y ; x = (m_1 - m_2) \div (f_1 - f_2)$$

In practice, the problem of combination arises because a third observation  $m_3$  referable to  $f_3$ , if paired off with  $m_1$  or  $m_2$ , will not yield exactly the same value of  $x$ . Accordingly, we regard any such measurement as subject to an *error*, which we denote as  $\epsilon_j$  for the  $j$ th of a sequence of such measurements, so that

$$\epsilon_j = (f_j \cdot x + y) - m_j \text{ or } m_j = f_j \cdot x + y - \epsilon_j \quad . \quad (i)$$

It is customary to speak of each such equation on the right as one of our *observational equations*. If we make  $n$  measurements we shall define the sum of the squares of the errors as

$$E = \sum_{j=1}^{j=n} \epsilon_j^2 \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

\* i.e. that the mean of a sufficiently large number of estimates based thereon is the true value we seek.



# METHOD OF LEAST SQUARES AND CONCEPT OF POINT ESTIMATION

The method of least squares prescribes that we shall choose as our values of  $x$  and of  $y$  those which minimise  $E$ , i.e. those for which

$$\frac{\partial E}{\partial x} = 0 = \frac{\partial E}{\partial y} \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

If  $\epsilon_j$  does *not*\* depend on the particular value of  $f_j$ , we may write:

$$\frac{\partial \epsilon_j^2}{\partial x} = 2f_j (f_j \cdot x + y - m_j) ; \quad \frac{\partial \epsilon_j^2}{\partial y} = 2 (f_j \cdot x + y - m_j)$$

Whence we obtain from (ii) and (iii)

$$\sum_1^n f_j (f_j \cdot x + y - m_j) = 0 = \sum_1^n (f_j \cdot x + y - m_j) \quad . \quad (iv)$$

We may write this result in the form

$$\sum_1^n f_j (f_j \cdot x + y) = \sum_1^n f_j \cdot m_j \quad . \quad . \quad . \quad (v)$$

$$\sum_1^n (f_j \cdot x + y) = \sum_1^n m_j \quad . \quad . \quad . \quad (vi)$$

We have thus two equations involving observed values of  $m_j$  and the unknown values of  $x$  and  $y$ . The numerical example given below exhibits the most convenient computing scheme for the solution; but an alternative way of writing down the solution will prove instructive at a later stage. The formal relation between the statistical procedure known as *regression* and the theory of the combination of observations will then be apparent if we proceed as follows. For brevity, we may write the mean values of  $f_j$  and  $m_j$  as  $M_f$  and  $M_m$ ; and

$$\sum f_j = nM_f ; \quad \sum m_j = nM_m ; \quad \sum f_j^2 = nQ_f ; \quad \sum f_j \cdot m_j = n \cdot C_{fm}$$

Whence we derive:

$$\begin{aligned} \sum (f_j - M_f)^2 &= nQ_f - nM_f^2 ; \\ \sum (f_j - M_f) (m_j - M_m) &= nC_{fm} - nM_f \cdot M_m \end{aligned}$$

Our equations are then

$$(n \cdot Q_f)x + (nM_f)y = nC_{fm} \text{ and } (n \cdot M_f)x + ny = nM_m$$

\* This assumption is amenable to experimental verification in the domain of physical measurement, but it is gratuitous in the domain of natural variation.

The solution for  $x$  is therefore :

$$x = \frac{C_{fm} - M_m M_f}{Q_f - M_f^2} = \frac{\sum (f_j - M_f) (m_j - M_m)}{\sum (f_j - M_f)^2} \quad . \quad (\text{vii})$$

More generally, for a series of measurements in linear relation to three unknown quantities, we may write our observation equations in the form :

$$m_j = a_j \cdot x + b_j \cdot y + c_j \cdot z - \epsilon_j$$

The reader will easily derive the least squares solution as that of the three equations

$$\left. \begin{aligned} \sum a_j (a_j \cdot x + b_j \cdot y + c_j \cdot z) &= \sum a_j \cdot m_j \\ \sum b_j (a_j \cdot x + b_j \cdot y + c_j \cdot z) &= \sum b_j \cdot m_j \\ \sum c_j (a_j \cdot x + b_j \cdot y + c_j \cdot z) &= \sum c_j \cdot m_j \end{aligned} \right\} \quad . \quad (\text{viii})$$

In this set-up, we need vary only the contributions of two unknown components by assigning different values to the corresponding constants (e.g.  $a_j$  and  $b_j$ ) so that we can fix the third (e.g.  $c_j = 1$ ). The method cannot yield more equations of the form shown in (viii) than the number of unknown quantities involved. Hence, it cannot lead to inconsistent estimates of the latter.

*Numerical Example:* If we here use data of a Wheatstone Bridge experiment of the type described above, as cited by Wald (*Theory of Errors and Least Squares*), to illustrate the computation in accordance with the pattern exhibited in (vi)–(viii), we should pause to ask ourselves whether we have not exceeded the terms of reference of the normal law of error. For we are now working with a null point instrument; and it is unlikely that repeated determinations of the total resistance of the circuit for the same fixed length of the bridge wire would transgress the boundaries of 5 scale divisions. The data of our experiment will be as follows for a bridge wire of 100 cm.

Length ( $l$ ) of Bridge Wire included (cm.)	Total Resistance (ohms.)
10	0.116
30	0.295
50	0.503
80	0.760

# METHOD OF LEAST SQUARES AND CONCEPT OF POINT ESTIMATION

Here our observations ( $r_{0,l}$ ) refer to the total resistance ( $r_{t,l}$ ) and we wish to estimate the resistance ( $u$ ) of the bridge wire alone; but to do this we have to eliminate the resistance ( $v$ ) of the external circuit. If  $l$  is the length of the bridge wire included in the circuit, the fraction of  $u$  which contributes to the total resistance is  $0.01l = a_l$ . Thus

$$r_{t,l} = a_l \cdot u + v$$

$$\epsilon_l = a_l \cdot u + v - r_{0,l}$$

The so-called normal equations prescribed by (iv) in this case will be:

$$\sum a_l(a_l \cdot u + v - r_{0,l}) = 0 = \sum (a_l \cdot u + v - r_{0,l})$$

The computation is as follows:

$$\begin{array}{rcl} 0.01u + 0.1v - 0.0116 = 0 & 0.1u + v - 0.116 = 0 \\ 0.09u + 0.3v - 0.0885 = 0 & 0.3u + v - 0.295 = 0 \\ 0.25u + 0.5v - 0.2515 = 0 & 0.5u + v - 0.503 = 0 \\ 0.64u + 0.8v - 0.6080 = 0 & 0.8u + v - 0.760 = 0 \end{array}$$

$$\text{Total } 0.99u + 1.7v - 0.9596 = 0 \quad 1.7u + 4v - 1.674 = 0$$

Thus our normal equations are:

$$0.99u + 1.7v - 0.9596 = 0$$

$$1.7u + 4v - 1.674 = 0$$

The solution is

$$u = 0.9277 ; v = 0.0243$$

If we employ the method of least squares we thus take 0.9277 ohms as our estimate of the resistance of the bridge wire itself.

*The Legendre-Laplace Axiom.* We have now acquainted ourselves with one class of problems, the earliest, in which it is customary to invoke the Method of Least Squares. At least we can here say that the river of theory has not overflowed its banks, i.e. our sole concern is with *error* in the Gaussian sense. We have, however, postponed the examination of the stochastic credentials of the criterion of preference (p. 188) which led the immediate predecessors of Gauss to endorse the method. We must now scrutinise it, because the Legendre-Laplace

Axiom, as we may call it, recurs in expositions of the method widely current throughout the nineteenth century; and it has emerged in a new guise during the past generation. First, it will be instructive to cite Merriman (*Introduction to Geodetic Surveying*, 1893):

The Method of Least Squares sets forth the processes by which the *most probable values of observed quantities* are derived from the observations. The foundation of the method is the following principle:

In observations of equal precision the most probable values of observed quantities are those that render the sum of the squares of the residual errors a minimum.

This principle was first enunciated by Legendre in 1805, and has since been universally accepted and used as the basis of the science of the adjustment of observations. The proof of the principle from the theory of mathematical probability requires more space than can be here given, and the plan will be adopted of taking it for granted. Indeed some writers have regarded the principle as axiomatic. (*Italics inserted.*)

This citation is of two-fold interest. First, it emphasises a minor postulate which is of major consequence if we transport the algebra of the theory of error into the domain of natural variation, where equal precision connotes the highly debatable assumption of equal dispersion of observed values about a suppositious norm. What is more relevant to our present purpose is that the obscurity of the idiom serves to conceal the irrelevance of anything in the classical theory to the postulate definitive of our best choice of a parameter value, viz. that (if true) it would assign maximum probability to the observed set of observations.

This irrelevance becomes apparent when we explore its implications in the set-up of a 2-stage classical model. We postulate five types of urns, A, B, C, D, E, the total number being  $N$  and the numbers of each type being  $N_a, N_b$ , etc., respectively. Thus we assign prior probabilities  $P_a = N_a \div N, P_b = N_b \div N$ , etc., to the initial act of choice. We shall refer to the number of red balls in a 4-fold replacement sample from one such urn as our observed quantity (*score*), and shall postulate  $p_a = 0, p_b = \frac{1}{4}, p_c = \frac{1}{2}, p_d = \frac{3}{4}, p_e = 1$  as the proportions of red balls

# METHOD OF LEAST SQUARES AND CONCEPT OF POINT ESTIMATION

in the several types of urn. Below we see the 4-fold sample distributions for each type of urn.

Score	0	1	2	3	4	Total
A	256	0	0	0	0	256
B	81	108	54	12	1	256
C	16	64	96	64	16	256
D	1	12	54	108	81	256
E	0	0	0	0	256	256
Total	354	184	204	184	354	1280

If we adopt the Legendre-Laplace axiom, the rule of conduct we shall consistently pursue is as follows:

If the score is 0, say that  $p = 0$   
 „ „ 1, „ „  $p = \frac{1}{4}$   
 „ „ 2, „ „  $p = \frac{1}{2}$   
 „ „ 3, „ „  $p = \frac{3}{4}$   
 „ „ 4, „ „  $p = 1$

If we postulate that there are five urns in all, one of each type, so that the prior probabilities are equal, we may tabulate as follows the probability of correctly asserting that a given hypothesis is correct on the basis of the evidence supplied by the four-fold sample score in accordance with the prescribed rule.

Score	$p = 0$	$p = \frac{1}{4}$	$p = \frac{1}{2}$	$p = \frac{3}{4}$	$p = 1$
0	$\frac{256}{354}$	0	0	0	0
1	$\frac{81}{354}$	$\frac{108}{184}$	$\frac{54}{204}$	$\frac{12}{184}$	$\frac{1}{354}$
2	$\frac{16}{354}$	$\frac{64}{184}$	$\frac{96}{204}$	$\frac{64}{184}$	$\frac{16}{354}$
3	$\frac{1}{354}$	$\frac{12}{184}$	$\frac{54}{204}$	$\frac{108}{184}$	$\frac{81}{354}$
4	0	0	0	0	$\frac{256}{354}$

In this table the diagonal terms downwards from left to right assign the probability of correct identification for any given value of the score. In each column the highest term lies on this diagonal; but the highest term may be appreciably less than a half. In accordance with the viewpoint of Laplace,

i.e. against the background of a Bayes model w.r.t. which the relevant prior probabilities are equal, the rule of procedure implicit in the axiom thus means: (a) the probability of choosing the correct hypothesis is greater than that of choosing any single alternative to it; (b) the probability of choosing the correct hypothesis is not necessarily greater than the probability of choosing a wrong one.

The second conclusion suffices to exclude the axiom from any formidable claim to consideration as a useful tool of inductive reasoning. The preceding one might have claims to consideration as a mere convention *en rapport* with the interpretation suggested at the end of this chapter (p. 206), if it were possible to justify it without invoking a factually irrelevant 2-stage model upholstered with an entirely arbitrary postulate inconsistent with the Forward Look. One example suffices to show that it has no general validity unless we do so. Without disclosing the trick to the player, we shall now add to our system another urn of type C ( $p = \frac{1}{2}$ ), and we may set out the arithmetic of the sampling system in the following schema:

Score	0	1	2	3	4	Total
A	256	0	0	0	0	256
B	81	108	54	12	1	256
C	16	64	96	64	16	256
C	16	64	96	64	16	256
D	1	12	54	108	81	256
E	0	0	0	0	256	256
Total	370	248	300	248	370	1536

Only one column of this table need now concern us. When the 4-fold sample score is unity the probability that the sample will come from an urn of type B is  $\frac{108}{248}$ . That it will come from an urn of type C is  $\frac{64+64}{248} = \frac{128}{248}$ . Now the axiom prescribes that we shall say that the sample comes from an urn of type B whenever the score is unity; but classical theory prescribes that such 4-fold samples containing a single red ball will turn up more often from an urn of type C. Within the framework of a suppositious 2-stage model approach, the criterion of preference embodied in the Legendre-Laplace axiom thus entitles

us to frame a rule for the choice of a particular parameter with an uncertainty safeguard less than what we should properly assign to the operation of the rule for the identification of any single admissible alternative parameter value as the true one, if (and only if) we embrace the principle of insufficient reason as an act of faith.

Since we shall see that Gauss repudiated it, the fact that the Legendre-Laplace axiom is intelligible only if we invoke the notion of inverse probability current in the historic context of the two authors would be merely a matter of historic interest, were it also true that the method of least squares is the only accepted recipe for *point-estimation*, i.e. for specifying some unique value of a parameter as the best one on the basis of information supplied by a finite sample of observations. The method of least squares is not in fact a unique prescription for doing so. It is one of special applicability in the domain of *measurement* (representative scoring). An alternative procedure with more relevance to the domain of *enumeration* (taxonomic scoring) adopted in our last model is the *Method of Maximum Likelihood* which relies on the same postulate, as we shall now see. The formal statement of the *Method of Maximum Likelihood* in no way involves the assumption of a normal—or of any particular—distribution law, a circumstance which may account for its appeal to theoreticians, albeit this is equally true (*vide infra*) of the Method of Least Squares; but its author commits us to the same profession of faith as the Legendre-Laplace axiom. As Kendall (*Biometrika*, Vol. XXXVI, 1949) remarks:

it states that we are to proceed on the assumption that the most likely event has happened. Now, why?

Why, indeed, unless we invoke the doctrine of inverse probability?

THE METHOD OF MAXIMUM LIKELIHOOD. The simplest of all model situations will serve to illustrate the alternative method of point estimation referred to by Kendall in the passage cited. Our supposition is that we take at random one penny from a bag containing pennies all of which have not the same *bias*.\*

\* As defined by v. Mises, see however p. 460.

We toss it four times, and record the result. Our problem is to identify the penny taken as one of a particular class of pennies in the bag, specifiable as such by its *bias*.

If  $p$  is the probability that a penny turns up heads in a single trial the long run frequency ( $y$ ) distribution for the number ( $x$ ) of heads in a 4-fold trial is:

$x$	0	1	2	3	4
$y$	$q^4$	$4pq^3$	$6p^2q^2$	$4p^3q$	$p^4$

We may speak of  $b$  as the bias in favour of heads if we write  $p = (\frac{1}{2} + b)$ . The above then becomes

$x$	0	1	2
$y$	$\frac{(1-2b)^4}{2^4}$	$\frac{4(1-2b)^3(1+2b)}{2^4}$	$\frac{6(1-4b^2)^2}{2^4}$
$x$	3	4	
$y$	$\frac{4(1-2b)(1+2b)^3}{2^4}$	$\frac{(1+2b)^4}{2^4}$	

If therefore the observed score of a single 4-fold toss is 3, the unknown probability of the event is

$$p_{3.4} = \frac{(1-2b)(1+2b)^3}{4}$$

Evidently,  $p_{3.4}$  will have two minima, viz. zero probability, when  $b = +\frac{1}{2}$  or  $-\frac{1}{2}$ , i.e. when  $p = 0$  or  $p = 1$ , to get three heads and one tail being impossible on either assumption, and by tabulating  $p_{3.4} = 4p^3(1-p)$  for different values of  $p$  and  $b$  as in the table on page 199, we see that it has a maximum value at or in the neighbourhood of  $p = 0.75$ ,  $b = 0.25$ .

We can get the value of  $b$  corresponding to the *exact* maximum of  $p_{3.4}$  in the usual way by equating the first derivative to zero, i.e.:

$$\frac{dp_{3.4}}{db} = \frac{6(1+2b)^2(1-2b) - 2(1+2b)^3}{4} = 0$$

$$\therefore 6(1+2b)^2(1-2b) = 2(1+2b)^3$$

$$\therefore b = 0.25 \quad \text{and} \quad p = 0.75$$



# METHOD OF LEAST SQUARES AND CONCEPT OF POINT ESTIMATION

On the basis of the evidence, viz. a score of three heads in a single 4-fold toss, we speak of the particular value  $b = 0.25$  as the maximum likelihood point-estimate of the bias in favour of heads and of the particular value  $p = 0.75$  as the maximum likelihood point estimate of the probability of the coin turning

p	b	$p_{3.4} = 4p^3(1-p)$
0	-0.5	0
0.05	-0.45	0.0005
0.10	-0.40	0.0036
0.15	-0.35	0.011
0.20	-0.30	0.026
0.25	-0.25	0.047
0.30	-0.2	0.076
0.35	-0.15	0.111
0.40	-0.10	0.154
0.45	-0.05	0.200
0.50	0	0.250
0.55	+0.05	0.299
0.60	+0.10	0.346
0.65	+0.15	0.384
0.70	+0.20	0.412
0.75	+0.25	0.422
0.80	+0.30	0.410
0.85	+0.35	0.368
0.90	+0.40	0.292
0.95	+0.45	0.171
1.00	+0.50	0

up heads in a single trial. We may generalise the foregoing procedure for situations in which the score is  $x$  in an  $r$ -fold trial so that

$$p_{x,r} = 2^{-r} \cdot r_{(x)} (1 - 2b)^{r-x} (1 + 2b)^x$$

$$\frac{dp_{x,r}}{db} = 2^{-r} \cdot r_{(x)} 2x (1 + 2b)^{x-1} (1 - 2b)^{r-x} - 2^{-r} \cdot r_{(x)} 2 (r - x) (1 - 2b)^{r-x-1} (1 + 2b)^x$$

On equating the last expression to zero, we get:

$$b = \frac{2x - r}{2r} \quad \text{and} \quad p = \frac{x}{r}$$

For 4-fold sample scores of 0, 1, 2, 3, 4 in the model situation invoked to illustrate the implications of the Legendre–Laplace axiom, the maximum likelihood estimates of  $p$  will therefore be as there cited:  $p = 0$ ,  $p = \frac{1}{4}$ ,  $p = \frac{1}{2}$ ,  $p = \frac{3}{4}$ ,  $p = 1$ . What we have already had occasion to infer thus applies with equal force to the method of maximum likelihood. The axiom cited by Kendall is a meaningful rule of conduct against the background of the doctrine of inverse probability; but if we relinquish the latter, it is devoid of any intelligible justification.

*The Principle of Minimum Variance.* If neither the method of least squares nor any other method of point estimation such as that of maximum likelihood can guarantee that the risk of accepting the value assigned by it to the unknown quantity is less than the risk of accepting any other *single* value, still less that the risk is necessarily small, we have to face the question: is there any special reason for adopting any such procedure? Accordingly, we have now to examine an alternative interpretation of the method of least squares, and one which does *not* invoke the Legendre–Laplace axiom. It is in fact that of Gauss with whose name later generations have come to associate the method of least squares, although current literature on the theory of errors still largely proceeds from assumptions which Gauss repudiated.

Misunderstanding concerning what Gauss proved is easy to understand. He wrote his *magnum opus* (1821) on the method in Latin, and in classical Latin at that. Bertrand's translation did not appear till many years after expositors of the principle had associated his name with the earlier views of Laplace; and it is still customary to give credit to Markhof (1912) for a theorem which Gauss proved three-quarters of a century before him.\* As stated, this theorem does not invoke the Legendre–

\* In an informative *Historical Note* on the Method, R. L. Plackett (*Biometrika*, 1949) states: "Gauss was the first who justified least squares as giving those *linear estimates which are unbiased of minimum variance* . . . (he) presented his justification in 1821. The paper is written in Latin, but a French translation was published by

Laplace axiom or any considerations which necessarily rely on the Bayes' scholium in the background. It does not indeed invoke a normal distribution of errors. It merely asserts that the method of least squares leads to an *unbiased estimate whose sampling variance is minimal*. The advantages of choosing an estimate with this end in view still raise controversial issues which Gauss could scarcely foresee; and we shall leave the discussion of them till a later stage. First, let us acquaint ourselves with what Gauss did indeed establish.

To justify the assertion that the method of least squares leads to an unbiased estimate of minimal variance when we use it to assign a value to component measurements of which we vary at least one as in the numerical illustration following (viii) above, it will suffice to disclose the pattern of the proof. Accordingly, we shall confine our attention to such a simple situation in which we wish to assign a value to a fixed component  $x$ , i.e. the resistance of the bridge wire on the basis of direct measurements  $m$ , depending on  $x$  and a second fixed component  $y$ . In the absence of any error, we postulate the theoretical relation

$$m_r = k_r x + y \quad . \quad . \quad . \quad . \quad . \quad (ix)$$

In this expression we are free to vary  $k_r$  on the assumption that any error involved in recording its value (the fraction of the length of the bridge wire) appears in our observation equations as a quantity independent of  $k_r$ . The observation equations then take the form

$$m_r = k_r x + y - \epsilon_r \quad . \quad . \quad . \quad . \quad . \quad (x)$$

Our observation equations will be consistent if we have only two values of  $m_r$ , and we need postulate only three values to exhibit the principle under discussion, i.e.

$$m_1 = k_1 x + y - \epsilon_1 ; m_2 = k_2 x + y - \epsilon_2 ;$$

$$m_3 = k_3 x + y - \epsilon_3 \quad (xi)$$

Our aim will first be to show how we can obtain an unbiased estimate ( $\bar{X}$ ) of  $x$  having minimal sampling variance by Bertrand in 1855 and the fundamental theorem incorporated in Bertrand's own book (*Calcul des Probabilités*) of 1888 . . . Gauss's proof is valid for all values of  $n$ , entirely free from any assumption of normality."

suitably weighting our observations ( $m_1$ , etc.). Accordingly, we write\* :

$$X = C_1 m_1 + C_2 m_2 + C_3 m_3 \quad . \quad . \quad . \quad (xii)$$

If our estimate is to be unbiased  $E(X) = x$ , so that

$$E(C_1 m_1 + C_2 m_2 + C_3 m_3) = x$$

$$\therefore C_1 \cdot E(m_1) + C_2 \cdot E(m_2) + C_3 \cdot E(m_3) = x$$

We are assuming that the errors cancel out in the long run, i.e. :

$$E(m_i) = k_i x + y$$

$$\therefore C_1 (k_1 x + y) + C_2 (k_2 x + y) + C_3 (k_3 x + y) = x$$

$$x = (C_1 k_1 + C_2 k_2 + C_3 k_3)x + (C_1 + C_2 + C_3)y$$

To satisfy this relation and hence to ensure that  $X$  in (xii) is an unbiased estimate :

$$(C_1 k_1 + C_2 k_2 + C_3 k_3) = 1 \text{ and } (C_1 + C_2 + C_3) = 0 \quad (xiii)$$

From (xiii) we obtain :

$$C_3 = \frac{1 - C_1 k_1 + C_1 k_2}{k_3 - k_2} ; C_2 = \frac{1 - C_1 k_1 + C_1 k_3}{k_2 - k_3} \quad (xiv)$$

The last equation defines the condition that  $X$  is an unbiased estimate of  $x$ . To say that it has minimal variance is to say that  $V_x = E(X - x)^2$  is as small as possible. Now the variance  $\sigma_i^2$  of the distribution of  $m_i$  depends only on the error distribution assumed to be independent of the particular value of  $k_i$ , i.e. :

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$$

If we adopt (xii) as our definition of  $X$  :

$$V_x = C_1^2 \sigma_1^2 + C_2^2 \sigma_2^2 + C_3^2 \sigma_3^2 = (C_1^2 + C_2^2 + C_3^2) \sigma^2 \quad (xv)$$

\* We here assume that the estimate  $X$  is to be a linear function of  $m_1, m_2$ , etc. This assumption calls for comment. One may argue as follows. Let  $X = f(m_1, m_2, \dots)$ , any continuous function of our observations, i.e. such that we can expand it as a power series, w.r.t.  $m_1, m_2$ , etc. By hypothesis the mean of first powers of the error terms will vanish, but the mean of even powers of the error terms will be positive. Hence  $X$  will be an unbiased estimate of  $x$  only if expressible in first powers of  $m_1, m_2$ , etc.

Having defined  $X$  in this way, we wish to specify  $C_1, C_3$ , etc., so that  $V_x$  is a minimum, i.e.

$$\frac{\partial V_x}{\partial C_1} = \frac{\partial V_x}{\partial C_2} = \frac{\partial V_x}{\partial C_3} = 0 \quad . \quad . \quad . \quad (xvi)$$

In this set of equations:

$$\frac{\partial V_x}{\partial C_1} = \left( \frac{\partial C_1^2}{\partial C_1} + \frac{\partial C_2^2}{\partial C_1} + \frac{\partial C_3^2}{\partial C_1} \right) \sigma^2$$

The condition of minimum variance for the choice of  $C_1$  is therefore

$$\begin{aligned} \frac{\partial C_1^2}{\partial C_1} + \frac{\partial C_2^2}{\partial C_1} + \frac{\partial C_3^2}{\partial C_1} = 0 &= 2C_1 + 2C_2 \frac{\partial C_2}{\partial C_1} + 2C_3 \frac{\partial C_3}{\partial C_1} \\ \therefore C_1 + C_2 \frac{\partial C_2}{\partial C_1} + C_3 \frac{\partial C_3}{\partial C_1} &= 0 \quad . \quad . \quad (xvii) \end{aligned}$$

To satisfy (xiv), we must put

$$\begin{aligned} C_2 \frac{\partial C_2}{\partial C_1} &= \frac{(1 - C_1 k_1 + C_1 k_3)(k_3 - k_1)}{(k_2 - k_3)^2} \quad \text{and} \\ C_3 \frac{\partial C_3}{\partial C_1} &= \frac{(1 - C_1 k_1 + C_1 k_2)(k_2 - k_1)}{(k_3 - k_2)^2} \end{aligned}$$

Whence (xvii) becomes:

$$\begin{aligned} (k_2 - k_3)^2 C_1 + (k_3 - k_1)(1 - C_1 k_1 + C_1 k_3) \\ + (k_2 - k_1)(1 - C_1 k_1 + C_1 k_2) = 0 \\ \therefore 2(k_1^2 + k_2^2 + k_3^2 - k_1 k_2 - k_1 k_3 - k_2 k_3) C_1 = 2k_1 - k_2 - k_3 \\ . \quad . \quad . \quad . \quad (xviii) \end{aligned}$$

We can simplify (xviii) by introducing  $M_k$  the mean value of  $k_r$ , i.e.:

$$3M_k = (k_1 + k_2 + k_3) \text{ so that } 3(k_1 - M_k) = 2k_1 - k_2 - k_3 \quad . \quad . \quad . \quad (xix)$$

$$\begin{aligned} 9(k_1 - M_k)^2 &= 4k_1^2 + k_2^2 + k_3^2 - 4k_1 k_2 - 4k_1 k_3 + 2k_2 k_3 \\ 9(k_2 - M_k)^2 &= k_1^2 + 4k_2^2 + k_3^2 - 4k_1 k_2 + 2k_1 k_3 - 4k_2 k_3 \\ 9(k_3 - M_k)^2 &= k_1^2 + k_2^2 + 4k_3^2 + 2k_1 k_2 - 4k_1 k_3 - 4k_2 k_3 \end{aligned}$$

$$\therefore \sum_{r=1}^{r=3} (k_r - M_k)^2 = 2(k_1^2 + k_2^2 + k_3^2 - k_1 k_2 - k_1 k_3 - k_2 k_3) \quad (xx)$$

By substitution from (xx) and (xix) in (xviii) we now obtain

$$C_1 = \frac{k_1 - M_k}{\sum_{r=1}^{r=3} (k_r - M_k)^2} \quad . \quad . \quad . \quad . \quad (xxi)$$

From the build-up of the equations, we can write down by inspection the corresponding identities which satisfy (xiii) and (xiv) as

$$C_2 = \frac{k_2 - M_k}{\sum_{r=1}^{r=3} (k_r - M_k)^2} \quad \text{and} \quad C_3 = \frac{k_3 - M_k}{\sum_{r=1}^{r=3} (k_r - M_k)^2} \quad (xxii)$$

By substitution of these weights in (xii) we thus obtain

$$X = \frac{\sum_{r=1}^{r=3} m_r (k_r - M_k)}{\sum_{r=1}^{r=3} (k_r - M_k)^2}$$

We may express this relation in another form. If  $M_m$  is the sample mean value of  $m_r$ :

$$3M_m = \sum_{r=1}^{r=3} m_r \quad \text{and}$$

$$(m_r - M_m)(k_r - M_k) = m_r(k_r - M_k) - M_m k_r + M_m M_k$$

$$\therefore \sum_{r=1}^{r=3} (m_r - M_m)(k_r - M_k)$$

$$= \sum_{r=1}^{r=3} m_r(k_r - M_k) - M_m \sum_{r=1}^{r=3} k_r + 3M_m M_k$$

$$\therefore \sum_{r=1}^{r=3} (m_r - M_m)(k_r - M_k) = \sum_{r=1}^{r=3} m_r(k_r - M_k) \quad (xxiii)$$

Whence we may write our unbiased estimate of minimal variance for  $n$  observational equations in the alternative form:

$$X = \frac{\sum_{r=1}^{r=n} (m_r - M_m)(k_r - M_k)}{\sum_{r=1}^{r=n} (k_r - M_k)^2} \quad . \quad . \quad (xxiv)$$

\*

\*

\*

\*

The foregoing derivation sufficiently illustrates the Gaussian rationale of the Method of Least Squares for combinations of observations in geodesy, astronomy and many situations which arise in the physical laboratory. Of more relevance to its uses in later statistical theory is the situation which arises when we wish to determine one or more physical constants. This will be the topic of our next chapter. Here we may pause with profit to recall the peculiar circumstances which impelled Gauss to confer on the theory of error subsequently associated with his name the prestige of a mathematician of such outstanding originality and foresight.

There is a topical piquancy in the association of his name with that of Bradley (p. 161) in this setting. With Bradley's (1728) observations on  $\gamma$  *Draconis* and with his own interpretation we now associate the discovery of a phenomenon provoking a new synthesis of theoretical optics and Galilean mechanics. From this fusion we may trace the origin of the dilemma which Lorentz and Einstein have successfully resolved to the satisfaction of their own contemporaries; and we think of Gauss first and foremost in the context of the theory of relativity as a pioneer of Non-Euclidean geometry. Gauss himself did indeed entertain a hope that later research seems to have vindicated. He turned to Bradley's careful and extensive observations on the stars Altair and Procyon for evidence of the inadequacy of the relevant Euclidean postulate with full appreciation that a true discrepancy would not have hitherto defeated recognition, unless of an order commensurate with the error of observation. Since it was not his primary concern to disclose a universal pattern for the combination of observations, still less to enunciate the basic principles of a calculus of judgments, we shall not denigrate his genius if we now ask what indeed was the outcome of an undertaking with no reward for his own expectations in his own time.

*Point Estimation as a Stochastic Procedure.* In this chapter, we have explored two different ways in which we may seek justification for the combination of observations by the Method of Least Squares. One invokes an axiom. As such, it is not susceptible of proof and its implications disclose no intelligible merits of the Method at an operational level, if we reject the

doctrine of inverse probability. We have also seen that the same axiom intrudes into the prescription of an alternative and lately more fashionable procedure for point estimation, i.e. for assigning a unique value of a metric or parameter\* on the basis of a pool of data from which it is possible to extract different estimates thereof. So far we have not asked what advantages the alternative approach confers. Before doing so, it is pertinent to remark that point-estimates derived by the method of maximum likelihood satisfy the requirement of minimum variance, if they also satisfy Fisher's criterion of sufficiency (p. 446), i.e. if a sufficient statistic exists.

To get the operational content of the question last stated into focus, let us recall the historical situation in which the theory of error took shape. If our concern, like that of the astronomer or of the surveyor, is to make a map of the heavens for the navigator or a map of a territory for the railway engineer, we shall need to specify each of a set of points uniquely. Thus some standard convention which everyone observes is a social discipline essential to geodesic or to astronomical enquiry. If content to look at the procedure of point estimation unpretentiously as a social undertaking, we may therefore state our criterion of preference for a method of agreement so conceived in the following terms:

(i) different observers make at different times observations of one and the same thing by one and the same method;

(ii) individual sets of observations so conceived are independent samples of possible observations consistent with a framework of competence, and as such we may tentatively conceptualise the performance of successive sets as a stochastic process;

(iii) we shall then prefer any method of combining

\* In geodesics we may illustrate the use of the Method to estimate by an indirect procedure which necessarily relies on more than one observation, an unknown distance, itself at least conceivably amenable to estimation by recourse to a single observation. In the next chapter our concern will be with the use of the Method to estimate the constant of a physical law deemed to be true on the basis of prior experience. By its definition, such a constant (here referred to as a *parameter*) is not amenable to direct observation.



constituents of observations, if it is such as to ensure a higher probability of agreement between successive sets, as the size of the sample enlarges in accordance with the assumption that we should thereby reach the true value of the unknown quantity in the limit;

(iv) for a given sample size, we shall also prefer a method of combination which guarantees minimum dispersion of values obtainable by different observers within the framework of (i) above.

In the long run, the convention last stated guarantees that there will be minimum disagreement between the observations of different observers, if they all pursue the same rule consistently. On the same understanding, the previous requirement specified by (iii) guarantees a procedure that will lead them to agree correctly, if they carry out a sufficiently large number of observations. In stating our programme of action in terms thus consistent with the Forward Look, we invoke no considerations of inverse probability; but we have brought into focus two issues which call for additional comment.

First, we have undertaken to operate within a fixed framework of repetition. This is an assumption which is intelligible in the domain of surveying, of astronomy or of experimental physics. How far it is meaningful in the domain of biology and whether it is ever meaningful in the domain of the social sciences are questions which we cannot lightly dismiss by the emotive appeal of the success or usefulness of statistical methods in the observatory, in the physical laboratory and in the cartographer's office. Aside from this, the foregoing considerations suggest a rationale of the Method of Least Squares *en rapport* with the Gaussian approach only if we identify minimum dispersion with minimum variance. Now variance is a unique measure of dispersion if we postulate a normal distribution of errors; but it is easy to construct distributions of which this statement is not true, and variance has then no special claims to commend it in preference to other measures of dispersion. Thus we have achieved little by renouncing the assumption (see footnote p. 200) that the Gaussian Law of Error is necessarily applicable to situations in which the

preference for an unbiased estimate of minimal variance is interpretable as a useful social convention.

If we do invoke the normal curve as a plausible description of the distribution of repeated observations subject to accidental (p. 205) errors alone, we expose ourselves to a temptation to err from the straight and narrow path of the classical theory of risks at a different level. Can we assign an uncertainty safeguard to the statement that an estimate  $x_m$  of the true value  $x_i$  lies within an interval  $x_i \pm c$ ? This is the form of a class of questions which are the theme of Chapter 18 on interval estimation. We shall there see reason to doubt that the specification of an unbiased point estimate of minimal sampling variance fulfils all the prescribed conditions; and if it does not, the disciplinary approach to point estimation is seemingly the only one consistent with the classical theory of risks.

If we adopt a consistently behaviourist approach, the terms of reference of a rationale of point estimation conceived simply in the foregoing terms as a social discipline essential to the art of map-making, celestial or terrestrial alike, encompass only the class of situations in which it is necessary to *locate putatively real points for further reference in a stable situation*. This is not the intention when we invoke the Method of Least Squares in anthropometric and social studies to prescribe fitting curves passing as nearly as may be through points referable to no assumed *true* values unless we identify the latter with some hypothetical non-existent norm. Our next task will therefore be to assess the relevance, if any, of the Gaussian theory of error to the current statistical procedure subsumed by the terms Regression, Multivariate Analysis and Analysis of Covariance.

The last remarks do not signify that point estimation is pointless in domains of enquiry other than astronomy, cartography and physics. In one branch of biological science its use is strictly on all fours with its use in surveying. No less than a map of Kentucky, a chromosome map of a species embodies information for future use in a definable framework of repetition; and the application of the Method of Maximum Likelihood by R. A. Fisher to the combination of observations which contribute to its construction accounts in no small measure

for the disposition of biologists to condone its use in enquiries which endorse no intention of future application at the level of numerical tabulation and no opportunities for using such tabulated information in comparable situations.

The view here stated is somewhat less iconoclastic than that of so eminent a theoretical physicist as the late Norman Campbell (1928):

I reject, then, the Gaussian theory of error, without qualification and with the utmost possible emphasis; and with it go all theoretical grounds for adopting the rules that are based on it. But the rules themselves are not necessarily worthless; confidence in them may be restored if some less fragile support is found. (*Measurement and Calculation*, p. 162.)

## CHAPTER NINE

# ERROR, VARIATION AND NATURAL LAW

IN CHAPTER ONE we have deferred the obligation to define sharply the terms of reference of the type of statistical procedures there called the *Calculus of Exploration*. As now seen, it emerges in the same historic *milieu* as the Calculus of Errors; when new precision instruments were available for use in astronomy and geodesics; but the end in view was wholly different. Ostensibly it was to fashion a new instrument for the discovery of scientific laws pertaining to human society. Its parent was Quetelet, by training an astronomer and a student of geodesics. He it was who first drew attention to the similarity between: (a) certain empirical distributions of variation w.r.t. measurable characteristics of individual members of a population; (b) the Gaussian distribution of instrumental errors in an observatory. Quetelet himself did not invite the ridicule of Bertrand and of other contemporary mathematicians by exploiting the algebraic opportunities of a metaphor with so luxuriant a subsequent overgrowth of portentous symbolism. The adaptation of the formal theory of the combination of instrumental observations to the study of individual variation in nature and in society was the outcome of the partnership between his disciples Francis Galton and Karl Pearson.

By the Calculus of Exploration in our own setting I here refer to the types of statistical procedure respectively called *regression* or *multivariate analysis* on the one hand and *factor analysis* on the other. The statistical concept common to both is *covariance*, a name which describes the mean of the products of the deviations of paired measurements or numbers (e.g. load and stretch of spring) from a putative true mean or from a non-existent norm (e.g. that of birth weights of babies and duration of pregnancy). As such, the mean value of the numerator of (xxiv) in Chapter 8 or in (iii) below defines its sample value. In this chapter and in the next we shall deal with regression only. Its formal relation to the theory of error is more obvious than that of factor analysis; and it brings more readily

into focus the additional assumptions we must invoke when we transgress from the proper domain of observational error into the uncharted terrain of natural variation.

We have now familiarised ourselves with the use of the method of least squares to determine the true value of a measurement in a situation comparable to those in which the problem of combining observations most commonly arises, i.e. in astronomy and in geodesics. Throughout the nineteenth century such was indeed the principal domain of its application. From a formal viewpoint, the procedure is precisely equivalent to a use of the method with supposedly far wider terms of reference, as when the end in view is to determine the slope constant of a simple linear physical law involving only two variable measurements. Hooke's law of the spring (*ut tensio sic vis*) will suffice to illustrate such a situation. That the law is linear does not restrict its interest unduly from the viewpoint of the laboratory worker, who commonly seeks a suitable formula by choosing a score transformation to give a good linear fit. Thus we can investigate Boyle's law ( $pv = k$ ) relating the volume ( $v$ ) to the pressure ( $p$ ) of a gas at fixed temperature by plotting pressure against density ( $d$ ) in accordance with the substitution  $kv^{-1} = Kd$  so that  $p = Kd$ .

If  $x$  is the weight applied and  $y$  is the length of the spring the customary school textbook statement of the law of the spring for the range in which it holds good is:

$$y = k.x + C$$

More precisely, what we mean is that the result of a sufficiently large number of determinations of  $y$  for one and the same value of  $x$  will yield a mean ( $M_{y,x}$ ) which satisfies the linear relation

$$M_{y,x} = k.x + C \quad . \quad . \quad . \quad . \quad (i)$$

That the value of the constant  $k$  depends on the dimensions and material of the spring we may make explicit by labelling it as  $k_s$ . When the end in view is to determine  $k_s$  for a particular spring, we may regard the weight ( $x$ ) in the scale-pan as subject to no error of observation, if we make repetitive

determinations of the length ( $y$ ) referable to one and the same value of  $x$ . More fully then, our *observational* equation referable to the  $j$ th measurement of  $y$  for a fixed ( $i$ th) value of  $x$  is:

$$y_{ij,s} = k_s \cdot x_{i,s} + C - \epsilon_{ij,s} \quad . \quad . \quad . \quad . \quad (ii)$$

In this expression we have merely interchanged our specification of  $k_s$  and  $x$  in (ix) of Chapter Seven as constants respectively referable to a particular experiment and to all experiments undertaken in situations to which the only sources of error are *accidental* in a sense to be defined more precisely at a later stage. In accordance with (xxiv) in Chapter Eight, the method of least squares prescribes as our estimate of  $k_s$  based on  $n$  paired values of  $x$  and  $y$ , each referable to a different value of  $x$ :

$$k_s = \frac{\sum_{i=1}^{i=n} (y_{i,s} - M_{y,s}) (x_{i,s} - M_{x,s})}{\sum_{i=1}^{i=n} (x_{i,s} - M_{x,s})^2} \quad . \quad . \quad (iii)$$

As stated above, we speak of the *mean value* of the products in the numerator of (iii) as the covariance of  $x$  and  $y$ , written  $Cov(x, y)$ . If  $m$  values of  $y$  are available for each of  $n$  values of  $x$ , so that the number of paired values is  $nm$  we must adjust (iii) accordingly, viz.:

$$k_s = \frac{\sum_{i=1}^{i=n} \sum_{j=1}^{j=m} (y_{ij,s} - M_{ij,s}) (x_{i,s} - M_{x,s})}{m \sum_{i=1}^{i=n} (x_{i,s} - M_{x,s})^2} \quad . \quad . \quad (iv)$$

In current textbooks of statistics for research workers in biology or in the social sciences it is customary to write (iv) in the form:

$$k_{xy} = \frac{Cov(x, y)}{V_x}$$

*Numerical Illustration.* To keep a foothold on the solid earth of the Gaussian domain, we may use the numerical data of a high school class experiment on the determination of the elastic constant within

# ERROR, VARIATION AND NATURAL LAW

the range of tension prescribed by Hooke's law to illustrate the foregoing use of the method of least squares:

<i>Load (grams)</i>	<i>Mean stretch (mm.)</i>
$x_i$	$M_i$
1	0.4
2	1.1
3	1.4
4	2.1

Any two of these pairs of values will suffice to yield an estimate of  $k$ , but we can derive six different estimates of  $k_{xy}$  by taking different pairs of paired values alone. To combine our observations with a view to choosing the unbiased estimate of minimal variance, we proceed as follows:

$$M_x = \frac{1}{4}(1 + 2 + 3 + 4) = 2.5$$

$$M_y = \frac{1}{4}(0.4 + 1.1 + 1.4 + 2.1) = 1.25$$

$$V_x = \frac{1}{4}(1 + 4 + 9 + 16) - (2.5)^2 = 1.25$$

$$\begin{aligned} \text{Cov}(x, y) &= \frac{1(0.4) + 2(1.1) + 3(1.4) + 4(2.1)}{4} \\ &\quad - (1.25)(2.5) = 0.675 \\ k_{xy} &= \frac{0.675}{1.25} = 0.54 \end{aligned}$$

This illustration of the use of the method raises no new issues in the context of the Gaussian theory. What gives it special interest is that it emerges in the Galton-Pearson partnership in a factually new setting. The *elastic* coefficient of the foregoing example is now grandiloquently the *Coefficient of Regression* of  $y$  on  $x$ . As such, it is the keystone of the imposing edifice currently referred to as multivariate analysis. For this reason, a comprehensive examination of the credentials of the Calculus of Exploration would be incomplete without a preliminary re-examination of the factual assumptions which justify the use of the Method of Least Squares to estimate a linear constant in the domain of physical experiment. Such is the theme of what follows. In Chapter Ten, I shall discuss the meaning of regression from a more formal viewpoint with the help of relevant stochastic models.

If we are to do justice to our theme, three issues will invite clarification :

- (i) what do we mean by *error* in this context ;
- (ii) what do we imply when we say that  $k_s$  is a constant *definitive of the law* as applied to a particular spring ;
- (iii) within what *framework of repetition* do we conceive a stochastic model to be relevant to our specification of  $k_s$  ?

*The Meaning of Error.* I cannot convey to the reader a correct view of the use of the term *accidental error* in a preceding paragraph better than by quoting at length from a book in the hope that any reader with erroneous ideas about the proper terms of reference of the Gaussian Theory of Errors will read it. In his authoritative text on *The Combination of Observations*, Brunt (1917) defines two classes of errors of observation as follows :

As we shall frequently have to refer to constant and systematic errors in the sequel, it will be well to have a clear conception of the meaning of these terms. A *constant error* is one which has the same effect upon all the observations in a series. It has the same magnitude and sign in all observations. A *systematic error* is one whose sign and magnitude bear a fixed relation to one or more of the conditions of observation. It should be noted that neither of these types of error fulfils the laws of accidental errors given. . . .

Errors of the types here referred to are not therefore errors which fall within the province of the theory of combination of observations. We assume that the investigator knows his job sufficiently well to eliminate all known sources of constant or of systematic error so defined, in particular to control external conditions to the best of his ability, to exclude personal bias and to make the appropriate correction for instrumental defects. The classical theory of Error does not take over till the research worker has shouldered this responsibility. Brunt defines its proper domain as follows (*italics inserted*) :

If a series of observations be made, and corrected in each case for the errors due to the three factors\* considered above, it will in

\* (Instrument, observer, external conditions.)



general be found that the corrected measurements differ among themselves. These individual differences are ascribed to a fourth class of error, known as the accidental error.

*Accidental errors* are due to no known cause of systematic or constant error. They are irregular, and more or less unavoidable. The term "accidental" is not used here in its ordinary significance of "chance." Strictly speaking, an observation of any kind is affected by the state of the whole universe at the time of observation. But as an observer cannot take account of the whole universe and its changes of condition during the time occupied by his observations, he has to limit his attention to those operative causes which he regards as affecting his observations in a measureable degree; i.e. he limits his attention to the "essential conditions." If an observation could be repeated a number of times, and corrected in each case for changes in the essential conditions, the results of all the observations should be identical. But in practice it is found that the individual observations in a series differ among themselves. These differences may be ascribed to the fact that the so-called "essential conditions" do not include all the effective operative causes. There will be other operative causes of error, whose laws of action are unknown, or too complex to be investigated. These causes will introduce errors which will appear to the observer to be accidental.

. . . when all the systematic errors traceable to the instrument, the external conditions, or the observer, have been corrected, no observation can be regarded as perfect. It will *miss perfection on account of the presence of accidental errors*. The effect of the accidental errors will differ for different observations in the same series. It is thus impossible to attain certainty in the result of an observation. In practice a series of observations is made in the hope that the discussion of the series will eliminate the effect of the accidental errors. The problem which we have to solve is that of deciding the best method of conducting this discussion. . . . In what follows, accidental errors will be regarded as obeying the following laws:

- (i) A large number of very small accidental errors are present in any observation.
- (ii) A positive error and an equal negative error are equally frequent.
- (iii) The total error cannot exceed a certain reasonably small amount.
- (iv) The probability of a small error is greater than the probability of a large error.

On a minor issue, I think that the last remarks are incomplete. In so far as any manageable theory of error invokes stochastic considerations, it implies a fifth postulate, i.e. that the errors are independent of one another and of the true value of the relevant metric. However, that is immaterial in this context. I have italicised one remark in the foregoing, because it has a special bearing on (iii) in the preceding statement of the postulated laws of accidental errors. Brunt's wholesome use of the term perfection confers no licence for the recurrently stated contemporary *credo* that statistics have freed science from the tyrannical chimera of absolute truth as the goal of enquiry. On the contrary, we rely on such a concept as the only tangible basis for a definition of what is erroneous, and the postulate mentioned makes this explicit. Why must we assign equal probability to negative and to positive errors? The answer is that the long-run mean error must be zero, if we are to endorse the possibility of approaching the true value as our accepted goal more closely as we enlarge the content of our experience.

Within the framework of the assumption that there is a true law, and that this true law would be specifiable in all relevant numerical particulars if our fallible methods of observation did not handicap our efforts, the concept of the mean is therefore something more than a convenient parameter of a stochastic distribution. *The mean of a sufficiently large number of observations on one and the same metric is the unique observation we are seeking imperfectly to record.* If this savours of metaphysics to the evangelists of the new doctrine of free stochastic grace, I am content to remark that they cannot have it both ways. If they appeal to the dominant role of statistical theory in contemporary science, they rest their case in no small measure on the reliance of the laboratory worker on statistical models in situations which can endorse a stochastic model with intelligible relevance if, and only if, we take truth seriously in the manner of our forefathers.

*The definitive parameter.* Let us now seek an answer to our second query: in what sense is  $k$ , a constant definitive of the law of a spring of specified materials and dimensions? In the practice of the laboratory, we assuredly presuppose what our examination of the foregoing question has brought into focus.

We do not invoke the Method of Least Squares, to establish the law. It comes into the picture *when we have already satisfied ourselves that the law is sound*. We start by comparing the outcome of repeated experiments which we perform with every possible precaution to eliminate (or correct for) personal, instrumental or environmental errors of the sort Brunt refers to as *constant* or *systematic*. We observe that each such experiment referable to the same range of  $x$  values yields a scatter of corresponding  $y$  co-ordinates on either side of a straight line. We likewise observe that the points on our scatter diagram do indeed cluster as closely round a straight line as our *prior knowledge* of the range of uncontrollable (*accidental*) error would entitle us to expect, if a straight line does indeed pass through the assumed *true values* of the relevant points whose co-ordinates  $(x_i, y_{ij})$  are definitive of a particular ( $i$ th) value of  $x$  and the corresponding  $j$  values of  $y$  referable thereto.

Having decided that a straight line is in this sense a satisfactory descriptive device for prescribing the result of stretching a spring in an assigned range of  $x$  values under carefully controlled conditions, we may then usefully incorporate our conviction in the corpus of scientific knowledge by tabulating for future use in appropriately defined situations the definitive parameter  $k$ , for springs of specified materials and dimensions, and thereby forestall unnecessary effort, e.g. when our aim is to prescribe the design of a spring balance. Only when we undertake this task of tabulation do we call on the Method of Least Squares, i.e. we do so to endorse what particular numerical values of  $k$ , appear in our *tables of physical constants*.

Whatever justification, if any, there may be for the Pearsonian reliance on the Method of Least Squares to define the straight line of best fit for the so-called linear regression of heights of brothers on heights of sisters, bodyweight of sons on bodyweight of mother or of family income on tuberculosis rates, it assuredly signalises an innovation on which the Gaussian theory of error confers no licence; and the more so if we invoke significance tests to justify our preference for a straight line rather than for one of an infinitude of other descriptive curves. If we then designate a *deviation* as the displacement of a point about some straight line around which our plotted points

appear to cluster, we most certainly repudiate the hope that any such line would go through every point if faultlessly determined. Faultless determination is indeed practicable in small scale undertakings involving enumerative data in both dimensions. In that event our deviations will certainly not be errors in the Gaussian sense of the term.

What then are such deviations which we find on our hands when we have indeed eliminated all error of observation? The answer is as inescapable as it is also simple. They record the failure of a straight line or some other descriptive fitting curve to tell us where a uniquely true value of  $y$  corresponding to a particular value of  $x$  lies, or vice versa. Their existence thus signalises the certainty that the selected descriptive curve is not a true law of nature as the astronomer or the experimental physicist interprets the term *law* in situations which provoke him to make use of the Method of Least Squares as a method of combining observations. The deviations we are discussing are not the outcome of human frailty. The points about which we are talking are not unique natural events that we are seeking to locate in our frail human way. The deviations are nature's errors, the failure of nature to get to the point. The point itself is a *norm* imposed on nature by the edict of a realm which is supernatural in any intelligible sense of the term.

In the grand manner of Bertrand (p. 172) we may answer the question proposed at the beginning of the last paragraph more picturesquely as follows. If we plot the tuberculosis rate against preassigned values of family income, each point the line goes through specifies a *normal* tuberculosis rate, not too tragic nor too encouraging to complacency. If we plot the heights of sisters against a fixed median height interval of the brothers, the point our line goes through specifies the height of the *normal* sister not too tall nor too short, not too domineering nor too clinging, not too glamorous nor too homely. If we plot the body weight of mothers against that of sons, the line goes through the body weight of the *normal* mother, neither too stout nor too skinny, neither too strict nor too indulgent, neither too flighty nor too immersed in domestic routine, not too dowdy nor too chic, neither passionate nor frigid. So we might continue with our catalogue of laws laid up in the

felicitous heaven of Plato's universals but devoid of conceivable content in the dismal terrain of the efforts of living men to come to grips with inexorable nature.

Whereas the practice of the ordnance survey, the observatory or the physical laboratory, relies on the Method to tabulate for *future use* without recourse to direct observation a parameter definitive of a law deemed to be the true one for reasons which have nothing to do with the method, too many who invoke it in the domain of anthropometry and social studies do so indeed to confer the status of law on a generalisation which is not a law of nature as laboratory workers speak of one. They do so, as we shall now see, in circumstances which preclude the possibility of ever making any practical use of the estimated parameter  $k$  prescribed within a definable framework of repetition. That such is at least a novel use of the Gaussian Theory of Error has not escaped the attention of writers on the uses of the theory within the framework of its original terms of reference. Thus Brunt remarks apropos its alleged utility in connexion with biological enquiry: "it is thus in no way justifiable to regard Least Squares as a magical instrument applicable to all problems."

*The Framework of Repetition.* We have seen how and why the laboratory worker, the astronomer or the surveyor may wish to determine a physical constant formally equivalent to what is in its latest reincarnation the *regression* coefficient, a neologism concocted to communicate Galton's erroneous beliefs about inheritance. The assumption is that we can define the circumstances in which an experiment or set of observations is *repeatable*. Otherwise the tabulation of values appropriate to particular statistics (e.g. elastic coefficients for springs of standard dimensions but different metals) would be valueless. All this presupposes that we can control the situation. It is implicit in what we here mean by saying that a law of experimental physics presumes a causal or, if we are too fastidious to use a word so *démodé*, a *consequential* relationship.

When we apply the Gaussian method of line fitting to measurements referable to concomitant variation of collateral relatives (e.g. heights of first cousins), it is obvious that no consequential relationship is conceivable. More generally in

sociology, we may hope or suspect that our data disclose a consequential relationship; but only controlled experiment in the most literal sense of both terms can disclose whether this is so or whether the relationship involved is *concurrent*, i.e. referable to a common antecedent. Perhaps Karl Pearson appreciated this uncertainty when he borrowed from Edgeworth (p. 178) a mathematical device which is meaningless in the setting of Gauss and has no bearing on the dilemma of geodesics. Partly because it is indeed so irrelevant to the combination of physical observations and partly because change of name by deed poll through publication of the main thesis in Series B of the *Transactions of the Royal Society* as a new rationale of Darwin's doctrine, the concept of Regression evaded the searchlight of Bertrand's critique. If sceptical about its practical value which they could assess, biologists were willing to concede its mathematical respectability on which they could not pass judgment with equally good grace. In such a setting, a new tribal ritual which derives no sanction from the admissible claims of point estimation in the world of affairs became fashionable by default.

The device we here recall is the *bivariate normal* universe. In a situation such as Hooke's law of the spring serves to illustrate, the investigator commonly records equal numbers ( $j$ ) of observations  $y_{ij}$  referable to each fixed value ( $x_i$ ) of the controlled variable, which does indeed then enjoy a special factual status in the formal statement of the law as the independent variable of the definitive equation. If we then postulate the normal law of accidental error, the universe of our sampling system conceived in stochastic terms against the background of a unique historical framework of repetition is a *normal-rectangular* universe. This convention is not inexorable; but in no thinkable circumstances consistent with the design of a physical experiment is it possible to conceive the chosen values of the fixed  $x$ -set as a sample taken randomwise from a normal universe of  $x$ -scores. To assert the contrary is indeed inconsistent with the identification of the independent variable with the variable subject to the control of the investigator.

Needless to say, brother's height has no priority w.r.t. sister's height or vice versa as the independent, i.e. controlled variable

of our scatter diagrams for regression of one on the other. To give verisimilitude to a framework of repetition prerequisite to a stochastic theory of sampling, and to accommodate the convenient fiction that height conforms to the normal law, we have thus to postulate a universe which is normal in two dimensions. The formal dilemma which arises out of this ambiguity of interpretation has recently prompted a statistician to propound the question: *are there two regressions?*\* It emerges because: (a) the choice of the so-called independent variable involves no *prima facie* priority when we plot two such variates as heights of brothers and sisters; (b) the same method leads to different values of the so-called regression coefficient if we substitute one for the other as the variable deemed to be independent.

The same dilemma does not arise in a controlled laboratory experiment. In plotting the results of such experiments we may distinguish between two procedures: (a) each value of the so-called *dependent* variate, e.g. the *stretch* of a spring, plotted against a particular value of the other (e.g. *tension* applied), may truly correspond to one value of the latter, as when we successively measure the stretch ( $y_{ij}$ ) produced by adding one and the same load ( $x_i$ ) to the scale-pan in a consistent environment; (b) each value of the dependent variate (e.g. *blood sugar*) plotted against one and the same value of the other variate (e.g. *insulin dosage*) involves an *unacknowledged* error of observation in the measurement of the latter. Either way, the customary procedure in the conduct of an experiment entails what we tacitly assume, as then formally admissible, when fitting a line to our observations by the method of least squares, viz. that all the errors of observation arise in assigning a value to the so-called dependent variate.

In the laboratory there is also commonly a clear-cut operational distinction between the variate we choose to designate as the dependent (e.g. *volume*) and the alternative one, i.e. the one which is more amenable to direct control (e.g. *pressure*). In applying the Gaussian method to laboratory data we may not indeed be free to make a choice between the two ways of fitting a line, though this is not always so. It may be possible

\* J. Berkson (1950). *Journ. Amer. Stat. Ass.*, 45, 164.

to adopt either of two procedures: (*a*) to measure the stimulus requisite to produce a fixed response; (*b*) to measure the response evoked by a fixed stimulus. In whichever way we do proceed, the value of the so-called independent variable may in fact be subject to experimental error, neglected as such by the way in which we plot our data. Customarily, the physicist allocates all sources of error to the side of the balance sheet identified with what he elects to call the dependent variate.

In laboratory enquiry, the mere fact that one variable is under the control of the investigator signifies that the relation sought is consequential; and the legitimate implications of the use of curve-fitting by least squares do not admit of any formidable dangers if we approach our task as an obligation to agree upon a social convention for reasons stated at the conclusion of Chapter Eight. That the method of least squares transplanted into the field of biology and sociology by Karl Pearson was indeed originally a device for use in the domain of a *theory of error*, thus imposes upon us the duty to scrutinise two basic assumptions we take for granted legitimately in laboratory practice: (*a*) the physicist can *control* every relevant variable in an experimental set-up other than variability of the type specified on p. 215 above; (*b*) all such deviations occur *random-wise* in repeated determinations.

In experimental biology, one must always and at all times also take stock of individual variation attributable to nature and to nurture; but the biologist with justifiable intention of propounding a law, as physicists use the term, e.g. the linear alignment of the genes, implicitly assumes the possibility of repeating observations based on different individuals without introducing a systematic source of variability referable to either. If he can specify the extent to which he has standardised the genetic make-up and culture of his stocks, there is then an intelligible framework of repetition within which the law is valid. It is scarcely in doubt that admissible postulates of physical experimentation and those the biologist may adopt with justifiable confidence on the understanding last stated are gratuitous in many situations which prompt sociologists and psychologists to employ regression equations.

In such enquiries, what is usually a more important source



of variation is a complex of external agencies we have no power to control. Were it otherwise, our residual variance, i.e. deviations from what our selected so-called law prescribes, would be simply a measure of the failure of our powers of observation to detect an explicitly definable and inflexible regularity of nature. As it is, our residual variance is, and to no small extent, a record both of the inadequacy of any simple law as a description of our observations purged of all error in the Gaussian sense and of our powerlessness to recreate a *unique historic event*. Thus the mortality experience of the Borough of Tottenham (Middlesex) 1952, summarised (inadequately) as a frequency distribution by age and income, defines a 3-dimensional unit sample distribution which has almost certainly never existed as a specification of the mortality experience of any pre-existing community and almost certainly can specify that of no community in the future. If we seek to formulate a rule of procedure in accord with the forward look, we have therefore no realisable prospect of opportunities for making any statements to which we can assign an upper limit of uncertainty consistent with the classical theory of risks.

Needless to say, this dilemma will not inhibit our industrious computations, if we have taken the decisive backward step which was the theme of Chapter Four, i.e. the identification of an ever-changing population of living creatures with a universe which remains constant within the indefinitely protracted framework of repetition prescribed by the classical theory. We can then locate in the Pearsonian manner, and in the Platonic heaven of universals, the 3-dimensional distribution of Tottenham mortality in the year 1952 as a sample of an infinite hypothetical trivariate normal universe from which (alas) only the hypothetical normal man with hypothetically normal opportunities presumptively enjoys the right to extract more than one sample.

To make the foregoing criterion more tangible, let us recall the law of the stretched spring. When we give assent to such a law, our presumptive aim is to state in advance how much we can extend a spring under specified loads, if we measure the extension with sufficient accuracy under specified conditions. A latent postulate is therefore that our laboratory is static,

since the results would be quite different in virtue of variations w.r.t. the gravitational constant  $g$ , if we made our observations in an aeroplane at different (and unknown) heights above sea-level. The best we could then hope for is that we could distribute our observations on the stretch with respect to a specified tension so that differences with respect to elevation would be uniformly distributed. Even in the absence of error inherent in the technique of observation as such, our line of best fit could then tally with the one definitive of the physical law of the static laboratory only in so far as it describes the trend of averages which have no definable bearing on future experience.

This parable gives us a new slant on the line which goes through the normal, as opposed to an *actual*, point of the Pearsonian regression graph. Figuratively speaking, the laboratory of the social scientist and of the vital statistician is most often an aeroplane at unknown and changing height above sea-level; and it is possible to formulate the implications of the foregoing remarks with reference to the assumed framework of repetition in which we conceptualise the sampling process at a more elementary level of discourse, if we recall a commonplace universally admitted as a canon of scientific method, viz. any statement of a scientific law is complete only in so far as it incorporates the recognition of its own limitations. To be sure, the laboratory worker familiar with such limitations may, and customarily does, forget to make them explicit; but he can commonly do so without compromising the usefulness of conclusions drawn from the law itself. For example, any bright boy who has passed through a high school course of physics can safely use an equation prescribing how the density of water varies in relation to temperature at sea-level pressure without succumbing to the temptation to invoke its aid to prescribe the density of steam at  $120^{\circ}\text{C}$ . and 730 mm. atmospheric pressure. Again, we all learn at school that Hooke's law breaks down, if the extension approaches breaking-point; and that van de Waals' equation must replace Boyle's simpler, and for many purposes good enough, rule in the neighbourhood of absolute zero or of the critical pressure.

That the *explicit* algebraic formulation of a physical law is always incomplete from this viewpoint is innocuous, because

the experimentalist translates it in the domain of action with the reservation that the correct interpretation carries with it a supplementary specification of the *boundary conditions* of its validity. To say this, is to say that the legitimate use of an equation definitive of a structural law in physics lies within the domain of *interpolation*, i.e. within the domain of a clearly conceived historic framework of repetition; and the teaching of elementary physics familiarises us with the absurdities which arise when we use such a law for *extrapolation* beyond the boundaries of its applicability. The teaching of sociology impresses the same lesson less firmly on the beginner. Extrapolation beyond its legitimate terms of reference is indeed precisely comparable to what we do, if we succumb to the temptation of using a regression equation as a basis for predicting how a wage increase will affect fertility or infantile mortality.

What is a sufficiently well-recognised truism in experimental science is a *caveat* we too easily ignore in sociology and vital statistics. We cannot legitimately infer from the regression of completed family size on family income what the completed family size would be, if we stabilised all incomes at a fixed level, thereby changing the framework of conditions in which the regression relation is valid. Statistical literature of the last fifty years abounds with generalisations of this sort; but it is not difficult to detect the fallacy in such reasoning, if we recall a fundamental difference between experimental investigation and statistical description, as already mentioned. We have previously had occasion to recognise that there is a clear-cut distinction in experimental science between what we commonly call the dependent and the independent, or as we might more informatively say, *consequent* and *antecedent*, variates. The antecedent (so-called independent) is the one which the investigator has under his direct and deliberate control; and commonly, though not always, it is the only one within his power to control. Thus one cannot fill an injection syringe with adrenaline solution by raising the blood pressure of the patient; but one can raise the blood pressure of the patient by injection of the contents of a syringe containing adrenaline in solution.

We recognise such a relationship as consequential because, and only because, we are able to interfere actively with the course of events; but we are not recording the result of any such active interference when we plot a regression graph of maternal morbidity on completed family size at a particular time in a particular place or of heights of first cousins referable to a particular and historically unique human population. At least as likely as not, the relationship involved is concurrent, and necessarily so w.r.t. the second example last mentioned. In any event, our plotted data can give us no assurance to the contrary. A little reflection on a simple suppositious situation will serve to clarify one set of limitations imposed on the use of statistical methods in sociological enquiry when we do indeed take the trouble to clarify the historical framework in which the enquiry itself proceeds.

We may imagine a set-up not uncommon in Asia or in Africa. A population subject to malaria lies spread over a dry hillside and over the swampy lowlands around it, the more prosperous house-holders having settled on the heights. In the nature of the case, we should then expect to find a correlation between mean income and malaria incidence in the various precincts. It would not then be surprising if we found that we could plot our statistics plausibly as a linear regression graph. In this situation, raising the income of the less prosperous sections of the community might admittedly permit more migration from the swampy lowlands and hence less risk of malaria; but only if there were still land available for building on the uplands and only if there were no commensurate increase in the value of house property. In the absence of any information about the availability of alternative accommodation and about the prospects of the building market, we therefore lack sufficient reason for inferring what effect an all-round increase of income would have. A vigorous planning policy to make available inadequately utilised housing accommodation might well produce beneficial results; but our regression equation contains no precise information relevant to this possibility. In any case, it cannot legitimately lead us to forecast the effects of a change defined uniquely in terms of the only relevant variable, i.e. income as such.

THE TYRANNY OF AVERAGES. The technique of simple regression suffices to illustrate all the essential features of multivariate analysis. In common with factor analysis, its initial programme is the description of populations of organisms in terms which exclude a prescription for the *control* of individual behaviour. Figuratively speaking, we might say much the same about the kinetic theory of gases; but the comparison would be superficial. At an operational level, the individual molecule, atom, electron and the like is merely a convenience for interpreting the behaviour of matter in bulk. In so far as we may speak of a gas as a population of molecules, it is the special task of the experimental physicist to teach us how to handle such populations with the assurance that we can define conditions relevant to the identification of other such populations alike in all relevant particulars. In large measure, the problems of greatest practical interest to the biologist and to the psychologist are problems of individual behaviour, and the end in view is to define conditions relevant to the identification of circumstances in which different individuals behave. The only proven method of achieving this is the Baconian recipe, i.e. successive elimination of relevant variables by punctilious regard for the canons of controlled experimentation.

Such is the issue which the great physiologist Claude Bernard, provoked by the influence of Gavarret (p. 97) and other of Quetelet's disciples inspired with the new evangel of averages, raises in one of his lectures posthumously published as *Introduction to the Study of Experimental Medicine*. I quote him at length, because the experimentalist is at least entitled to the last word on what class of problems he prefers to investigate:

By destroying the biological character of phenomena, the use of *averages* in physiology and medicine usually gives only apparent accuracy to the results. From our point of view, we may distinguish between several kinds of averages: physical averages, chemical averages and physiological and pathological averages. If, for instance, we observe the number of pulsations and the degree of blood pressure by means of the oscillations of a manometer throughout one day, and if we take the average of all our figures to get the true or average blood pressure and to learn the true or average number of pulsations, we shall simply have wrong numbers. In

fact, the pulse decreases in number and intensity when we are fasting and increases during digestion or under different influences of movement and rest; all the biological characteristics of the phenomena disappear in the average. Chemical averages are also often used. If we collect a man's urine to analyse the average, we get an analysis of a urine which simply does not exist; for urine, when fasting, is different from urine during digestion. A startling instance of this kind was invented by a physiologist who took urine from a railroad station urinal where people of all nations passed, and who believed he could thus present an analysis of *average* European urine! Aside from physical and chemical, there are physiological averages, or what we might call average descriptions of phenomena, which are even more false. Let me assume that a physician collects a great many individual observations of a disease and that he makes an average description of symptoms observed in the individual cases; he will thus have a description that will never be matched in nature. So in physiology, we must never make average descriptions of experiments, because the true relations of phenomena disappear in the average; when dealing with complex and variable experiments, we must study their various circumstances, and then present our most perfect experiment as a type, which, however, still stands for true facts. In the cases just considered, averages must therefore be rejected, because they confuse, while aiming to unify, and distort while aiming to simplify. Averages are applicable only to reducing very slightly varying numerical data about clearly defined and *absolutely simple cases*. . . . I acknowledge my inability to understand why results taken from statistics are called *laws*: for in my opinion scientific law can be based only on certainty, on absolute determinism, not on probability. . . . In every science, we must recognise two classes of phenomena, first, those whose cause is already defined; next, those whose cause is still undefined. With phenomena whose cause is defined, statistics have nothing to do; they would even be absurd. As soon as the circumstances of an experiment are well known, we stop gathering statistics: we should not gather cases to learn how often water is made of oxygen and hydrogen; or when cutting the sciatic nerve, to learn how often the muscles to which it leads will be paralysed. The effect will occur always without exception, because the cause of the phenomena is accurately defined. . . . Certain experimenters, as we shall later see, have published experiments by which they found that the anterior spinal roots are insensitive; other experimenters have published experiments by which they found that the same roots were sensitive. These cases seemed as comparable as

possible; here was the same operation done by the same method on the same spinal roots. Should we therefore have counted the positive and negative cases and said: the law is that anterior roots are sensitive, for instance, 25 times out of a 100? Or should we have admitted, according to the theory called the law of large numbers, that in an immense number of experiments we should find the roots equally often sensitive and insensitive? Such statistics would be ridiculous, for there is a reason for the roots being insensitive and another reason for their being sensitive; this reason had to be defined; I looked for it, and I found it; so that we can now say: the spinal roots are always sensitive in given conditions, and always insensitive in other equally definite conditions.

How far statistical methods can indeed contribute to the initial stage of an investigation by screening likely clues is not the issue which Claude Bernard raises in this context. In contemporary terms, what he rightly rejects is the now widely current practice of publishing as a discovery what is at best an encouragement to further examination. If his comments on the use of averages were salutary in his own time, they are still more so in ours. In his day, the statistician was content to assert a modest claim to the supervision of interpretation. In ours, he claims the prerogative of prescribing the *Design of Experiments*, i.e. the conduct of experiments to supply figures which certain test and estimation procedures can, or purport to be able to, accommodate. Indeed one contemporary publication,\* referring to the use of factor analysis in physiological enquiry, goes so far as to say:

Although it is true that any experimental statistical design permits one to look for a law in the relation between certain variables, without entering the experiment with too definite a hypothesis as to what form the law must take, *factor analysis is unique in demanding no prior hypothesis and in being automatically productive of a hypothesis.* (*Italics inserted.*)

Though the acceptance of such claims is the repudiation of the methods by which the experimental sciences have attained their present status, no experimental scientist of Bernard's standing has hitherto accepted the challenge publicly. It is

\*Cattell, R. B., and Williams, N. F. V. M. (1953). *Brit. Journ. Prev. Soc. Med.*, Vol. VII.

therefore pertinent to state that the issue is essentially a matter of scientific method on which mathematicians as such have no prior claim to arbitrate. If statistics is indeed the science of averages, the statistician has a special claim to our attention when the end in view is to record averages, as is true of demographic studies and administrative enquiries which do not concern themselves with individuals as such; but it is for the biologist, for the psychologist and for the clinician themselves to decide whether an average can be a satisfactory answer to questions they ask about the individual organism. The statement that a solution of iodine in KI makes starch paste blue in  $75 \pm 1.5$  per cent of samples examined will not commend itself to a chemist. He will wish to know what impurities or what range of pH values, etc., determine when the reaction does or does not occur. Such has been the attitude in which physiologists have hitherto undertaken the investigation of animal behaviour. If we abandon it, we are lowering our standards. We are concealing our retreat from a position of hard-won advantage behind an impressive smoke screen of irrelevant algebra.

What commonly evades recognition when the statistician advances such claims is that terms such as *law* and *hypothesis* have a totally different connotation in the operational domain of experimental enquiry and in the descriptive domain of computing indices referable to natural populations or to the variable states of single individuals. This will become more apparent in the next two chapters when we have examined both the origin of the term regression against the background of Galton's attempt to create a science of genetics without recourse to the safeguards of controlled experimentation and the subsequent contribution of factor analysis to our present knowledge of human behaviour. Meanwhile, the reader needs no intimate knowledge of the theoretical basis of the P-technique to appreciate what the writers last cited do in fact mean by a *hypothesis*. In the summary of Cattell and Williams we find that:

The correlation of 36 physiological variables and nineteen psychological variables measured on a normal 23-year-old male



for 110 successive days has yielded clusters of significant correlation coefficients which indicate that a considerable fraction of the day-to-day variation of measurement is due to fluctuation of unitary underlying functions rather than experimental error of measurement. . . .

In general, these results show that factor analysis is capable of structuring a wide array of physiological manifestations in a way not possible by any other method. *Controlled experiments, however, are desirable, if these findings by the exploratory, factor-analytic method are to be followed up to give causal explanation of the associations.* They also show substantial relations between the physiological patterns and some of the factor patterns previously recognised in psychological (total behaviour) variables. It is to be hoped that physiologists specialising in particular fields may be able to suggest more detailed explanations for some of these observed correlations. (*Italics inserted.*)

Thus the outcome is to *classify* certain phenomena which may prove to be worthy of the attention of the laboratory worker; but this is admittedly useful only in so far as it furnishes the latter with a clue. In what sense then do the writers use the term *hypothesis*? As I see it, they do so in the sense that the systematic botanist might so speak of the choice of the number of ridges on the ovary of an *Umbellifer* as a marker-characteristic to sort out genera embracing species with many common characteristics. That such taxonomical industry has its uses, no experimental biologist would deny; but classificatory systems as such have no intrinsic validity. All one can say about a taxonomy is that it is more or less useful as a means of identifying entities with common characteristics or as an impetus to investigate otherwise unrecognised similarities. Since no recipe for classification can anticipate what class of similarities will prove to be a fruitful theme for experimental enquiry, the use of the word *hypothesis* to signify a taxonomical convention is admissible only if we recognise that the word has a totally different meaning in the context of the laboratory.

STOCHASTIC MODELS FOR SIMPLE  
REGRESSION

IN WHAT has gone before I have assumed that the reader understands the contemporary connotation of the term *regression*, in so far as it is a procedure prescribed for fitting a line—not necessarily straight—to a scatter diagram of points defining paired score-values plotted on a 2-way Cartesian grid—or more generally a specification of the definitive parameters of a formula involving sets of  $n$  corresponding connected score-values. Thus we may choose to plot in a 3-dimensional grid family income, completed size of family and mortality rates as an assemblage of 3 corresponding connected score values. We have seen that the prescribed procedure, including the use of the familiar covariance formula for simple linear regression involving only two such variates, enlists the method of least squares originally invoked for use in situations essentially different from those which prompt sociologists and agricultural scientists to rely on it. We have also acquainted ourselves with the social setting in which the Method of Least Squares gained a footing in the laboratory at a time when the normal curve attained a peculiar status in descriptive statistics of populations; but we have not as yet examined the circumstances in which the Gaussian concept of point estimation first intruded into the domain of descriptive statistics.

It will help us to get into focus a distinction emphasised in Chapter Eight and to view it from a fresh angle, if we now retrace our steps to the historical background of the innovation signalled by Pearson's partnership with Galton. It may also be profitable to do so for another reason. We shall be better able to see how the tyranny of averages checked the progress of biological science at a time when new experimental methods (*vide* Chapter XIII) had laid the sure foundations for spectacular advances on a wide front. First, it will be necessary to explain how the definitive constant of a linear physical law such as the elastic coefficient of a spring has acquired so singular a designa-

tion as the *regression coefficient*, the more so since at least nine contemporary biologists out of ten, if under 50 years of age, and perhaps 99 sociologists out of a hundred, appear to believe that regression signifies the comparatively recent introduction of an algebraic manipulation peculiar to the realm of biological and/or sociological investigation. Nor are there many sources to enlighten those who cannot recall, as can the writer, the intellectual climate of the college or campus in 1912, when lectures on what we now call genetics impressed on the beginner that there are two sorts of heredity, severally designated alternating, qualitative or Mendelian and statistical, quantitative or biometric.

In another context I have referred to Galton's *Natural Inheritance* as source material for the genesis of the Quetelet *mystique*. Here our concern is to elucidate the origin of the word *regression* which makes its first appearance therein. Like Quetelet, Galton collected data w.r.t. such human measurements as heights, but with a different end in view. He industriously made scatter diagrams exhibiting values referable to pairs of individuals at a particular level of family relationship. For instance, he tabulated heights of sons against those of all fathers whose heights lie within a certain interval, e.g. 0.5 in. Having done so, he recorded with a rapture commensurate with the ambiguity of the outcome that the mean heights of sons lie very close to a straight line. Seemingly, he had expected that the grand mean of all sons in such a set-up would be the same as that of all the corresponding fathers. It turned out to be nearer the mean of the population, whence he discerned a universal drift of the hereditary process to what we may faithfully call in his own words *mediocrity*.

Galton christened the drift *regression*, a name which has outlived its original connotation and now signalises any exploits of draughtsmanship directed to exhibit the mean of a set of measurements plotted against particular values of another connected set, more especially perhaps when the draughtsman himself is not quite clear about what he is doing. Galton himself thought about the matter more deeply; and his second thoughts accommodated the circumstance that boys also have mothers. Being a gentleman, he was willing to apportion the

blame equally. He allocated to each parent a 25 per cent contribution to their son's height with a joint contribution of 50 per cent to his mediocrity. It was thus that there came to him the revelation embodied in a generalisation which incorporates a felicitously elementary arithmetical operation already beatified in the doctrine of Malthus. If a quarter of our inheritance comes from our Victorian father and a quarter from the weaker vessel, each of our four grandparents must contribute an eighth. Thus we arrive by easy stages at the quaint geometric series released to an expectant world as the *Law of Ancestral Inheritance*. Galton summarises it in the following passage cited from *The Average Contribution of each several Ancestor to the total Heritage of the Offspring*. (*Proc. Roy. Soc.*, 1897, Vol. LXI, pp. 401-5):

The law to be verified may seem at first sight too artificial to be true, but a closer examination shows that prejudice arising from the cursory impression is unfounded. This subject will be alluded to again, in the meantime the law shall be stated. It is that the two parents contribute between them on the average one-half, or  $(0.5)$  of the total heritage of the offspring; the four grandparents, one quarter or  $(0.5)^2$ ; the eight great-grandparents, one-eighth, or  $(0.5)^3$ , and so on. Thus the sum of the ancestral contributions is expressed by the series  $(0.5) + (0.5)^2 + (0.5)^3$ , etc., which, being equal to 1, accounts for the whole heritage.

The same statement may be put into a different form, in which a parent, grandparent, etc., is spoken of without reference to sex, by saying that each parent contributes on an average one-quarter or  $(0.5)^2$ , each grandparent one-sixteenth, or  $(0.5)^4$ , and so on, and that generally the occupier of each ancestral place in the  $n$ th degree, whatever be the value of  $n$ , contributes  $(0.5)^{2n}$  of the heritage.

The law to be verified supposes all the ancestors to be known, or to be known for so many generations back that the effects of the unknown residue are too small for consideration. The amount of the residual effect, beyond any given generation, is easily determined by the fact that in the series  $\frac{1}{2} + \frac{1}{4} + \frac{1}{8}$ , etc., each term is equal to the sum of all its successors. Now in the two sets of cases to be dealt with the larger refers to only two generations, therefore as the effect of the second generation is  $\frac{1}{4}$ , that of the unknown residue is  $\frac{1}{4}$  also. The smaller set refers to three generations, leaving an unknown residual effect of  $\frac{1}{8}$ . These large residues cannot be

ignored, amounting, as they do, to 25 and 12·5 per cent respectively. We have, therefore, to determine fixed and reasonable rules by which they should be apportioned.

It will be convenient to use the following nomenclature in calculations:

$a_0$  stands for a single member of the offspring.

$a_1$  for a single parent;  $a_2$  for a single grandparent, and so on, the suffix denoting the number of the generation. A parallel nomenclature, using capital letters, is:

$A_0$  stands for all the offspring of the same ancestry.

$A_1$  for the two parents;  $A_2$  for all the four grandparents, and so on. Consequently  $A_n$  contains  $2^n$  individuals, each of the form  $a_n$ , and  $A_n$  contributes  $(0\cdot5)^n$  to the heritage of each  $a_0$ ; while each  $a_n$  contributes  $(0\cdot5)^{2n}$  to it.

The analytical profundities of the foregoing train of reasoning do not take within their scope the possibility that the human family hands on a physical and cultural environment as well as an assemblage of what we now call *genes*. So we need not break a long since defunct butterfly on the wheel with the sledgehammer blows of a generation instructed in the art of conducting genetic experiments in standard culture conditions. It will suffice to recall the gratitude of the undergraduate of 1912 when his teachers abandoned this half of their assigned schedule with evident relief. The time had now come to talk of Mendelism, as they then said. Throughout the rest of the course one never lost the feeling that the prison gates were at last open. In those days, it was my own privilege to sit at the feet of Leonard Doncaster who discovered sex-linked inheritance, a phenomenon inconsistent with the half-and-half law briefly expounded above.

The vindication of Mendel in 1902 might have been the death-blow to the so-called law of ancestral inheritance, if the spiritual marriage of Galton and Pearson had not already produced a large progeny of mathematical contributions. As already stated, Pearson published these under a biological title in the *Philosophical Transactions of the Royal Society*, and on that account exempt from exposure to the scrutiny of mathematical colleagues sufficiently familiar with the theory of the combination of observations to recognise its abuse as well as its

uses. Pearson had already discerned in Galton's data a formal similarity between a graph exhibiting the several heights of sons of fathers of the same height and a graph exhibiting the several observations of the stretch of a spring when we apply the same load thereto. Accordingly, he proceeded to adapt the formal algebra of the nineteenth century theory of error to the requirements of an entirely new class of situations. Being a skilful manipulative mathematician equipped with a vigorous command of the English language, he had no difficulty in recruiting a militant following to spread a gospel which handicapped the progress of experimental genetics in Britain for at least half a generation.

Let us then be clear about what we mean by regression. When we plot pairs of measurements and discern a drift, regression is merely a name to signify what we discern. For the time being, we need not ask whither we are then drifting. If the drift suggests a straight line we speak of it as linear. More precisely, we may define *simple linear regression* as follows, in accordance with (i) of Chapter Nine. For  $j$  values of observations  $y_{ij}$  associated with a particular value of an associated measurement  $x_i$ , we may define a mean value  $M_{y,i}$ . If  $M_{y,i}$  increases by equal increments corresponding to equal increments of  $x_i$ , we may write  $M_{y,i} = k_{yx} \cdot x_i + C$ . We then say that linear regression is exact and speak of  $k_{yx}$  as the coefficient of regression of  $y$  on  $x$ .

*A Stochastic Model of the Gaussian Theory.* At one level we have seen that the nineteenth century theory of error is neither inconsistent with the eighteenth century theory of risks in games of chance nor necessarily cognate thereto. One can make different models such as the Hagen model to show that the normal is a good fitting curve for an empirical frequency distribution of repeated measurements of one and the same physical entity; but it is impossible to prove the validity of the postulates peculiar to one or other model and equally impossible to deny with a good grace the inadequacy of the normal law to describe all error distributions which arise in experimental enquiry.

At a different level, we may say that the Gaussian Theory invokes a stochastic model whose properties are unique. They

fully satisfy the requirements of the classical theory if we concede that a fixed randomwise distribution of accidental errors is consistent with the presumptive framework of repetition. This postulate of the theory is not amenable to conclusive demonstration; but it is not repugnant to common sense; and it derives sanction from laboratory experience as an at least plausible assessment of the outcome of repeated efforts to measure the same entity in comparable circumstances. To get into focus the affinity of the Theory of Error with the classical theory of risks in games of chance at this level, it will therefore be profitable to discuss a model game to which we may indeed prescribe the rules in accordance with the gentlemanly convention that both players have an equal chance of winning. The model appropriate to the Gaussian Theory will then help us to see how much factually relevant information we sacrifice on the altar of algebraic generalisation when we invoke such devices as the bivariate normal distribution to interpret or to misinterpret seemingly linear relationships in sociological enquiry.

The game we shall prescribe is one we may call the *Player-Bonus Model*. Player A tosses twice an ordinary cubical die and records his 2-fold score sum ( $x_a$ ). Player B then tosses 3 times a tetrahedral die as before with face scores — 1, 0, 0, + 1, adds the score sum ( $x_{b,0}$ ) of his own 3-fold toss to 3 times the score of player A at each trial and records as his gross score:

$$x_b = 3x_a + x_{b,0}$$

Player B wins if his mean score is greater than 3 times that of player A. Player A wins if his mean score is more than a third of that of player B.

By recourse to a simple chessboard diagram we obtain the frequency ( $f_a$ ) distribution of the score of player A as below:

*Distribution of A score*

$x_a$	2	3	4	5	6	7	8	9	10	11	12
$f_a \times 36$	1	2	3	4	5	6	5	4	3	2	1
Bonus ( $3x_a$ )											
of Player B	6	9	12	15	18	21	24	27	30	33	36

The distribution of the individual score of player B before addition of the trial-bonus is deducible in the same way as :

*Distribution of B's individual score*

$x_{b,0}$	- 3	- 2	- 1	0	+ 1	+ 2	+ 3
$f_{b,0} \times 64$	1	6	15	20	15	6	1

To construct a so-called correlation table which exhibits the relevant data of a scatter diagram referable to all possible samples and specifying their long run frequencies, we then proceed as follows. For A scores of 2, the distribution of B scores is that of  $6 + x_{b,0}$ , i.e. :

$x_b$	3	4	5	6	7	8	9
<i>Relative frequency</i>	1	6	15	20	15	6	1

Similarly for  $x_a = 3$  we have :

$x_b$	6	7	8	9	10	11	12
<i>Relative frequency</i>	1	6	15	20	15	6	1

To combine our entries of corresponding A-scores and B-scores with due regard to the distribution of the former, we must weight them accordingly by recourse to the relative frequencies of the A-scores themselves as given above. We then obtain the composite table shown opposite (page 239) :

Inspection of this table shows that  $M_{b,a}$ , the mean B-score for a fixed value of the A-score, increases by equal increments  $\Delta M_{b,a} = 3$  for unit change of  $x_a$  in the range  $x_a = 2$  to  $x_a = 12$  and  $M_{b,a} = 6$  to  $M_{b,a} = 36$ . If  $C = 0$  and  $k_{ba} = 3$ , we may therefore write

$$M_{b,a} = k_{ba} \cdot x_a + C$$

This is the type of relationship already defined as *exact linear regression*. Let us now look at the way in which it arises. We started with the relation :

$$x_b = 3x_a + x_{b,0}$$



# CHESSEBOARD DIAGRAM OF CONTINGENT A-SCORE AND B-SCORE

		A-score											
		2	3	4	5	6	7	8	9	10	11	12	
B-score	3	1	..	..									
	4	6	..	..									
	5	15	..	..									
	6	20	2	..									
	7	15	12	..									
	8	6	30	..									
	9	1	40	3	..								
	10	..	30	18	..								
	11	..	12	45	..								
	12	..	2	60	4	..							
	13	..	..	45	24	..							
	14	..	..	18	60	..							
	15	..	..	3	80	5	..						
	16	..	..	..	60	30	..						
	17	..	..	..	24	75	..						
	18	..	..	..	4	100	6	..					
	19	..	..	..	..	75	36	..					
	20	..	..	..	..	30	90	..					
	21	..	..	..	..	5	120	5	..				
	22	..	..	..	..	..	90	30	..				
	23	..	..	..	..	..	36	75	..				
	24	..	..	..	..	..	6	100	4	..			
	25						..	75	24	..			
	26						..	30	60	..	..	..	
	27						..	5	80	3	..	..	
	28						..	..	60	18	..	..	
	29						..	..	24	45	..	..	
	30						..	..	4	60	2	..	
	31							..	..	45	12	..	
	32							..	..	18	30	..	
	33							..	..	3	40	1	
	34								..	..	30	6	
	35								..	..	12	15	
	36								..	..	2	20	
	37									..	..	15	
	38									..	..	6	
	39									..	..	1	
Mean B-score		6	9	12	15	18	21	24	27	30	33	36	
(M <sub>b,a</sub> )													
Variance $\sigma^2_{b,a}$		1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	

# STATISTICAL THEORY

In the foregoing expression  $x_b$  signifies a single observation on the gross score of player B within the framework of a limitless game. In this game, the long-run mean value of  $x_{b,0}$  is zero for all values  $x_a$  may assume. The outcome is that the long-run mean value of the gross score of player B is three times that of player A in an indefinitely protracted sequence of trials specified by any particular score player A may record. The result would be just the same if the individual score contribution ( $x_{b,0}$ ) were zero at each trial, i.e. if the die of player B had zero score on each face.

Our parable is now complete. We may exhibit the correspondence between the score components of this set-up, and the variables of a physical experiment referable to a linear law involving 2 variables as below :

Stochastic Model	Symbol	Physical Experiment
Score of Player A.	$x_a$	Antecedent (independent) variable.
Trial bonus allocated to Player B.	$k_{ba} \cdot x_a$	True value of the consequent (dependent) variable.
Individual Score of Player B.	$x_{b,0}$	Error of observation.
Gross score of Player B.	$x_b = k_{ba} \cdot x_a + x_{b,0}$	Observed value of the consequent.
Mean value of the individual B-score for fixed value of the A-score.	$M_{b,0} = 0$	Zero mean of error distribution for each value of the antecedent.
Mean value of gross B-score for fixed value of the A-score.	$M_{b,a} = k_{ba} \cdot x_a$	True value of the consequent for fixed value of antecedent.

Needless to say, our final table of p. 239 does not summarise the results of a particular game. It summarises the long-run outcome of an endless sequence of games. In so far as our model is a model relevant to the assumptions inherent in the

Gaussian Theory of Error, our table is therefore a summary of a sample distribution referable to sampling randomwise from a bivariate universe of paired scores. In one respect, it is artificial, and inevitably so. No fixed rule prescribes the equivalent A-score distribution of repeated trials in the conduct of experimental enquiry. All that we can say about it is that it would never be normal, and perhaps there is too much of a soupçon of normality about the A-score distribution the rule of our game prescribes. It is at least unimodal and symmetrical, but a rectangular distribution of the A-scores as interpreted above is, as elsewhere stated, more consistent with the practice of the laboratory worker. However, this is immaterial to our undertaking. We can impose the condition last stated by a slight modification of the rule of the game, i.e. we may allow player A to toss the cubical die once only at each trial.

*A Modernised Stochastic Model of Galton's Regression.* In the foregoing exposition of a classical model which incorporates the essentials of the Theory of Error, we have initially stated the rules of a game of chance, and subsequently exhibited the correspondence between the formal prescription of the rules and the postulates of the Gaussian Theory in the domain of a linear physical law involving only two variables. Let us now change our tactics by: (a) first formulating the Pearson concept of linear regression in its original setting in terms of a scoring system; (b) next specifying the rules of a game consistent with the scoring system so prescribed.

As regards (a), we first recall that the term regression came into use to accommodate Galton's ill-starred superstitions about inheritance. So we may profitably choose the correlation of height of brother with height of sister. Since nothing human interference can do to a brother necessarily affects the height of a sister, we cannot speak of either variable in terms of antecedent and consequent. Thus the scatter diagram in this situation is necessarily referable to a *concurrent* relationship. Fortunately, we can approach our problem with information derived from a field of research which Pearson valiantly obstructed during his lifetime. We shall thus speak of different *genes* and different *environments* in which the same gene complex is consistent with unique manifestations of structure and/or

behaviour. Accordingly, we may schematically postulate component scores contributory to resemblance or "regression" as follows:

(i) *Contributory to resemblance.*

- $x_a$  the autosomal genes common to the gene complex of both sibs
- $x_m$  the equipment of *maternal* X-borne genes which brothers of the same parents share with their sisters
- $x_n$  the contribution attributable to environmental agencies (nurture) operating on both sibs in the same way.

(ii) *Contributory to so-called regression.*

- $x_{a.b}, x_{a.s}$  autosomal genes respectively peculiar to the gene complex of the brother and the sister
- $x_{m.b}, x_{m.s}$  *maternal* X-borne genes *ditto*
- $x_p$  the paternal X-borne gene equipment of the sister *only*
- $x_{n.b}, x_{n.s}$  the contribution of *differential* environmental agencies operating on the two sibs
- $x_{e.b}, x_{e.s}$  independent errors of measurement w.r.t. the observed characteristic, here assumed to be *height*.

\*

\*

\*

\*

The foregoing will suffice for our purpose, if we collect our data as did Galton. If we also include unlikesex twin pairs in our pool, we shall need to distinguish two other components of resemblances:  $x_u$  to specify what they owe to the fact that they share the same *uterine* environment at one and the same time and  $x_b$  to specify that they turn up in the same stage of the family fortunes by virtue of identical birth rank. If we include brothers and sisters singly or both adopted into different foster homes, we shall also need to distinguish common score components:  $x_f$  for sibs brought up together and  $x_s$  for sibs reared apart in homes at the same social level.

The refinements last mentioned need not detain us here. If we interpret the addition sign *with due safeguards* as below, we may specify the total paired scores of our scatter diagram, i.e.

the height of a boy ( $x_{h.b}$ ) and that of his sister ( $x_{h.s}$ ) by the equations:

$$x_{h.b} = (x_a + x_m + x_n) + (x_{a.b} + x_{m.b} + x_{n.b} + x_{e.b})$$

$$x_{h.s} = (x_a + x_m + x_n) + (x_{a.s} + x_{m.s} + x_{n.s} + x_p + x_{e.s})$$

In this context, the invocation of the genotypically normal sib conceived in a normal womb, reared on normal milk in a normal home will not assist us to interpret the assumed *additive* property in realistic biological terms consistent with what we now know. We are at liberty to consider: (a) each of the components  $x_{n.b}$ ,  $x_{n.s}$  as a (positive or negative) *increment* of height referable to the result of rearing the same pair of sibs in one and the same environment; (b) the components  $x_{a.b}$ ,  $x_{a.s}$ ,  $x_{m.b}$ ,  $x_{m.s}$  as increments referable to the substitution of a common gene complex for the entire assemblage of genes w.r.t. each of which one or both parents of a sib pair are heterozygous. Now no unique framework of substitution accommodates either class. Each such interpretation of the additive property embraces an infinitude of possibilities. Unless we dismiss the indisputable reality both of *interaction* between genes and of interaction between gene substitutions and specific environmental agencies, the effect of standardising either our environment or our gene complex in any one such way will determine what increment we can predicate of our common score components  $x_a$ ,  $x_m$ ,  $x_n$ .

Needless to say, Galton's approach to inheritance took no stock of such interaction. If we are to interpret correlation of relatives in his way, we shall therefore regard our paired total scores as the simple sum of two components  $x_c$  and  $x_{d.b}$  or  $x_{d.s}$  defined as follows:

$$x_c = x_a + x_m + x_n$$

$$x_{d.b} = x_{a.b} + x_{m.b} + x_{n.b} + x_{e.b}$$

$$x_{d.s} = x_{a.s} + x_{m.s} + x_p + x_{n.b} + x_{e.b}$$

Our equations definitive of score components now become:

$$x_{h.b} = x_c + x_{d.b}$$

$$x_{h.s} = x_c + x_{d.s}$$

On the assumption that we gratuitously (and erroneously)

interpret the additive sign in the usual way, we may speak of the above as *linear equations of concomitant variation*. With due regard to the reservations stated above, and neglecting interaction by assuming the independence of  $x_c$ ,  $x_{a,b}$ ,  $x_{a,s}$ , we may then define a stochastic model which precisely embodies the formal assumptions of our nature-nurture-error schema. It is the one I have called the *Umpire Bonus Model*. As a game which endorses the equal chance of success to each player, the prescription is as follows:

- (i) At each trial the players A and B each toss different dice, the number of tosses ( $r_a$  and  $r_b$ ) allocated being such as to guarantee the same mean value of the trial score sum in the long run, so that  $r_a = r_b$  if the two dice are alike;
- (ii) At each trial an umpire C tosses a third die  $r_c$  times, and each player records as his total score ( $x_a$  or  $x_b$ ) the score sum of his own toss ( $x_{a,0}$  or  $x_{b,0}$ ) added to that ( $x_c$ ) of C.

With this prescription, we can set out a table for the bivariate universe of the sampling process like that of p. 239, if we specify  $r_a$ ,  $r_b$ ,  $r_c$  and the types of dice respectively allocated to A, B and C. In the Pearsonian jargon we then say that:

- (i) there is linear regression of the B-score on the A-score ( $x_a$ ), if the mean B scores ( $M_{b,a}$ ) associated with particular values of the A-score increase by equal increments corresponding to equal increments of the latter ( $x_a$ );
- (ii) there is linear regression of the A-score on the B-score ( $x_b$ ) if the mean A-scores ( $M_{a,b}$ ) associated with particular values of  $x_b$  increase by equal increments corresponding to equal increments of  $x_b$  itself.

Before proceeding, we may usefully recall the true role of the umpire (C) in the game. The game itself is a game within the terms of reference of the classical theory of risks only if it is repeatable, i.e. relevant to the conduct of an experiment only in culture conditions of which we can predicate a distribution independent of and constant throughout a putatively randomwise sampling process. The umpire's score  $x_c$  is by definition what sibs of a pair owe both to a common ancestry and to a common environment. Hence it is not surprising that

samples\* of identical twins respectively reared in the same families and apart may (and do) yield different Pearsonian measures of correlation. Our umpire bonus is not a measure of genetic identity *per se* and has therefore no necessary bearing on the recognition of laws pertaining to inheritance. However, it is not our main concern in this context to conduct an autopsy on a defunct biological superstition. What is more relevant to the credentials of statistical procedures is why Galton or Pearson believed that any intelligible interpretation of the correlation of relatives implies a law of linear regression.

Elderton (*Frequency Curves and Correlation*) first propounded this model for a limited class of situations consistent with the specification I have given above, i.e. when all the dice are alike. It then happens that regression is linear in both dimensions as above, i.e.

$$M_{b.a} = k_{ba} \cdot x_a + C_b \quad \text{and} \quad M_{a.b} = k_{ab} \cdot x_b + C_a$$

If  $r_a = r_b$ , we shall find that  $k_{ba} = k_{ab}$ ; but otherwise there will be, as Berkson would say, two regressions. Now neither Elderton, who specifies this model in numerical terms without formal treatment of the postulates, nor Rietz† (1920) who later makes unnecessarily heavy weather of the algebra, appears to have noticed what is most instructive about its properties. If the dice are not identical, four situations may arise: linear regression in both dimensions of the scatter diagram, linear regression in one dimension only or in the other only and linear regression in *neither* dimension. That linear concomitant variation, as defined above for two concurrent variates, does not necessarily imply linear regression of either w.r.t. either, a single example serves to put beyond dispute.

The game we shall now play is in all respects like the one

\* In terms of the supplementary components specified above but neglected in the foregoing and with due regard to the limited meaning we here attach to the addition symbol, we may specify

Ordinary sibs reared together:	$x_c = x_a + x_m + x_f,$
„ „ apart, same social level:	$x_c = x_a + x_m + x_s$
„ „ apart, different social level:	$x_c = x_a + x_m$
Twins reared together	$x_c = x_a + x_m + x_u + x_f,$
„ „ apart, same social level:	$x_c = x_a + x_m + x_u + x_s,$
„ „ apart, different social level:	$x_c = x_a + x_m + x_u$

† *Ann. Maths.*, Vol. 21, 1920.

last prescribed. The only details to fill in are the specifications of the dice and numbers of tosses at each trial, viz. :

- (i) the umpire C tosses twice a flat unbiased circular die with 2 pips on one face and 1 pip only on the other;
- (ii) player A tosses once an unbiased tetrahedral die with face scores of 1, 2, 3 and 4 pips respectively;
- (iii) player B tosses once an unbiased tetrahedral die with 2 pips on each of three faces and 1 pip on the fourth.

We may proceed effortlessly as follows. Our grid of concomitant A-scores and B-scores before addition of the umpire's bonus is at each trial as below, the frequency of each bivariate cell score  $(x_{a,0}, x_{b,0})$  in that order being equal in accordance with the lay-out :

		A-score ( $x_{a,0}$ ) before addition of bonus			
		1	2	3	4
B-score ( $x_{b,0}$ ) before addition of bonus	1	1,1	2,1	3,1	4,1
	2	1,2	2,2	3,2	4,2
	2	1,2	2,2	3,2	4,2
	2	1,2	2,2	3,2	4,2

Our umpire bonus score ( $x_c$ ) distribution for the 2-fold toss is 2, 3, 3, 4. Accordingly, we construct 4 corresponding grids of bivariate scores as below :

$x_c = 2$				$x_c = 3$			
3,3	4,3	5,3	6,3	4,4	5,4	6,4	7,4
3,4	4,4	5,4	6,4	4,5	5,5	6,5	7,5
3,4	4,4	5,4	6,4	4,5	5,5	6,5	7,5
3,4	4,4	5,4	6,4	4,5	5,5	6,5	7,5
$x_c = 3$				$x_c = 4$			
4,4	5,4	6,4	7,4	5,5	6,5	7,5	8,5
4,5	5,5	6,5	7,5	5,6	6,6	7,6	8,6
4,5	5,5	6,5	7,5	5,6	6,6	7,6	8,6
4,5	5,5	6,5	7,5	5,6	6,6	7,6	8,6



We now condense our 64 entries in a frequency grid as follows:

		Total A-score ( $x_a$ )						$M_{a,b}$
		3	4	5	6	7	8	
Total B-score ( $x_b$ )	3	1	1	1	1	0	0	$\frac{1890}{420}$
	4	3	5	5	5	2	0	$\frac{2058}{420}$
	5	0	6	7	7	7	1	$\frac{2370}{420}$
	6	0	0	3	3	3	3	$\frac{2730}{420}$
$M_{b,a}$		$\frac{405}{108}$	$\frac{477}{108}$	$\frac{513}{108}$	$\frac{513}{108}$	$\frac{549}{108}$	$\frac{621}{108}$	

\*                      \*                      \*                      \*

Here it is neither true that  $M_{a,b}$  increases by equal increments when  $x_b$  also increases by equal increments, nor that  $M_{b,a}$  increases by equal increments when  $x_a$  does so. Even if our highly schematic interpretation of the correlation of brother's height with sister's height—front stall exhibit in Pearson's rationale of Galton's so-called Law of Ancestral Inheritance—is defensible in all particulars, we thus see that stochastic theory gives us no guarantee of linear regression unless we invoke some additional and exceptionable assumptions. This raises the question: what distribution of the component variates  $x_c$ ,  $x_{a,b}$ ,  $x_{a,s}$  in our schematic definitive equations of  $x_{h,b}$ ,  $x_{h,s}$  will ensure that linear regression is a necessary consequence of linear concomitant variation? Evelyn Fick (1945)\* has lately defined the necessary and sufficient condition in the domain of continuous variates. It appears that regression will then be linear if, and only if,  $x_c$ ,  $x_{a,b}$ ,  $x_{a,s}$  are normally distributed variates.

The consideration last stated makes Pearson's bivariate normal distribution an attractive playground for exhibiting a so-called Law of Inheritance which has not stood the test of

\* *Proc. Berkeley Symposium on Mathematical Statistics and Probability*. Edit. J. Neyman.

carefully controlled experimental studies of animal and plant breeding; but we have seen that the bivariate normal distribution does not—and cannot—describe a universe consistent with the proper terms of reference of the Theory of Error *sensu stricto*. Nor is it consistent with the considerations which led Pearson to propound his system of curve fitting by moments in a memoir ostensibly devoted to the theory of evolution. Why he chose this medium of publication is relevant to earlier remarks on the irrelevant invocation of the Central Limit Theorem, at least by implication, in Galton's own defence of his speculations and computations.

We have seen two reasons why this theorem is irrelevant to the contention that nature operates on stochastic principles applicable to an indefinitely large number of "contributory causes." We then have to assume that their effects on the growth process of a living being or of a society are independent and additive, a postulate which is indisputably false in a large variety of relevant situations. Furthermore, the endorsement of this supposition excludes the possibility of meeting skew distributions of characteristics of populations to which the theorem is supposedly relevant. Seemingly, Pearson did not appreciate the force of the first objection. Indeed, the superstition that we can rightly regard the effects of gene substitutions and of the multiplicity of interchangeable external agencies contributory to development as both independent and additive still persists, being inherent in Fisher's use of *Analysis of Variance* to assess the role of nature and nurture. What did profoundly disturb Karl Pearson is the fact that skew distributions do commonly occur in nature.

This is the theme of the latter part of (9) and the beginning of (11) in the second of the *Contributions to the Mathematical Theory of Evolution* with the sub-title *Skew Variation in Homogeneous Material*. It appeared in the Biological Series of the *Philosophical Transactions of the Royal Society* in the year 1895. The title and the content of the passages referred to explain what would otherwise be enigmatic. Why did the Pearson system of moment fitting curves see the light in a biological publication? Why does Pearson derive the parent differential equation of his Type system from considerations based on the

hypergeometric distribution, and why accordingly does a fitting curve for sampling without replacement from a finite binomial appear as head of the list as Type I?

A simple answer resolves these questions. If nature's urn of innumerable contributory small, and supposedly independent, causes whose effect is additive is also infinite, algebra offers us only the solution which the Central Limit Theorem endorses; but then we have to explain away the fact that empirical frequency distributions of population parameters are not necessarily—or even commonly—symmetrical. Happily it seems, we can sidestep this discouraging reflection by conceding that the contents of our nature-bag, if relevant to any particular make-believe sampling process congenial to the presumptions of Quetelet's *mystique*, is indeed finite and that the sample our supposititious supernatural gamester takes therefrom is a relatively large fraction of the whole.

It thus turns out that the bivariate normal frequency surface is not adequate to shoulder the burden of heredity on Galton's own terms; but the history of this engaging mathematical toy is of interest *vis-à-vis* the contemporary revaluation of statistical procedures from another viewpoint. When we formulate a stochastic model in the factual language of the discrete universe of scores which circumscribe the programme of the classical theory of risks in games of chance, we can make explicit the factual relevance of our algebraic conventions to the causal *nexus* which is the focus of interest to the naturalist and man of affairs. For instance, we can then see in what circumstances linear regressions may arise; and the examples cited above illustrate but two among many for which it is possible to specify an appropriate model game. When we transfer the issue to the continuous domain, our geometrical conventions give us no clue whatever to the diverse ways in which seemingly identical statistical patterns may arise in nature. A dominant *motif* of theoretical statistics for more than a century has been this disposition to sacrifice factual relevance to the dictates of mathematical tidiness.

As an index of resemblance between relatives, Pearson sidestepped the ambiguity of the two regressions by recourse to the *product-moment* coefficient of correlation. When regression

is indeed linear in both dimensions, this is the *geometric mean of the two coefficients*. By definition in accordance with (iv) of Chapter Nine

$$k_{ab} = \frac{\text{Cov}(x_a, x_b)}{V_b} ; k_{ba} = \frac{\text{Cov}(x_a, x_b)}{V_a}$$

As shown in Appendix 2, this is a tautology true of any set of paired scores which satisfy the condition that the row and column means of one set increase by equal steps for equal increments of the corresponding border scores of the alternative set. Whence we define the correlation coefficient ( $r_{ab}$ ) as

$$r_{ab} = \frac{\text{Cov}(x_a, x_b)}{\sqrt{V_a \cdot V_b}}$$

As a summarising index, the product moment coefficient has the useful property that its value is either  $+1$  (positive correlation) or  $-1$  (negative correlation) if there is one-to-one linear correspondence between  $x_a$  and  $x_b$ . When this is so,  $x_a = M_{a,b}$  for a fixed value of  $x_b$  and  $x_b = M_{b,a}$  for a fixed value of  $x_a$ . If  $M_{b,a}$  is constant for all values of  $x_a$  and  $M_{b,a}$  is constant for all values of  $x_b$  or if the row and column means ( $M_{b,a}$ ,  $M_{a,b}$ ) vary periodically and symmetrically with the border scores ( $x_a$ ,  $x_b$ ) its value is zero.

As also shown in Appendix 2, the fundamental property of the foregoing model when each player receives the same bonus from the umpire is that the variance ( $V_c$ ) of the umpire score distribution is equal to the covariance of the players' joint score distribution. If we write  $V_b$  and  $V_s$  respectively for the variances of the distribution of heights of brothers and heights of sisters, the product-moment coefficient for height on the foregoing arbitrary assumptions will thus be

$$r_{bs} = \frac{V_c}{\sqrt{V_b \cdot V_s}}$$

In deriving this expression, we assume that the score components  $x_c$ ,  $x_{d,b}$ ,  $x_{d,s}$  with sampling variances  $V_c$ ,  $V_{d,b}$ ,  $V_{d,s}$  are independent, so that

$$V_b = V_c + V_{d,b} \quad \text{and} \quad V_s = V_c + V_{d,s}$$

If regression is linear in both dimensions and  $k_{bs} \simeq k_{sb}$ , it follows from the definition of the regression coefficients that  $V_b \simeq V_s$  and  $r_{bs} \simeq V_c \div V_b$ , so that  $r_{bs} \simeq V_c \div (V_c + V_{a.b})$ . Pearson found that  $r_{bs} \simeq 0.5$ . This implies that  $V_{a.b} \simeq V_c \simeq V_{a.s}$ , if, as stated, the two coefficients are nearly equal; but there is no known reason why this should be so. Thus it is difficult to see how an appropriate stochastic model can endorse the arithmetic of the so-called law of ancestral inheritance. Still less is it easy to justify one assumption we have so far taken for granted. Our score components are conceptual entities about which we claim to know nothing more, in Galton's statement of the problem, than that  $x_c$  is referable to some assemblage of particles. By what sanction do we then claim that the components  $r_c$ ,  $x_{h.b}$ ,  $x_{h.s}$  are random variables? The only possible answer is a novel extension of the principle of insufficient reason. We have to assume randomness because we know nothing to the contrary. By the same token, a philosopher unacquainted with the railway timetable might admissibly assume that trains move in all directions random-wise throughout the course of the day.

The same gratuitous assumption is inherent in subsequent use of the product-moment coefficient as the keystone of the edifice called factor analysis. That it does indeed underlie much of contemporary thought gives the formal concept of the Irregular Kollektiv advanced by von Mises a special interest. If we take the contribution of von Mises seriously, we commit ourselves to the assertion that randomness is a *knowable* property of a system of scores. At least, it is a property whose existence we cannot conclusively demonstrate, but one we can infer with legitimate confidence against the background of large-scale experience. Thus the invocation of the concept in the domain of the statistical theory of the combination of instrumental observations is not on all fours with its intrusion in the domain of intangible hypothetical particles unless it leads to the construction of a unique hypothesis amenable to independent verification in the domain of experiment. The astronomer with long experience of his observatory may draw on a fund of factual information about how successive uncontrollable errors turn up in the course of his work; but the score components

which a stochastic model of regression or factor analysis (*vide infra*) must accommodate are fictions about whose distribution we can have no immediate experience.

We may now sum up in the following terms the outcome of our examination of the terms of reference of the theory of regression in this chapter and in its predecessor:

(1) The Gaussian calculus prescribes a discipline to promote agreement about the true values of physical measurements and of physical constants of laws inferred to be true for reasons to which the calculus is irrelevant; and the end in view is to place on record numerical values for future use in strictly comparable circumstances, endorsed either by experience of inexorable natural periodicities such as the movements of celestial bodies or by confirmatory experience of independent observations made daily in the laboratory or in the factory.

(2) The theory of regression prescribes a method of locating averages or specifying parameters of descriptive formulae referable to such averages; but the deviation of any individual score from such an average has no more title to be more or less near to a *true* value than any other; and the descriptive formula which embraces them can rightly claim the title of law as the physicist uses the term only if we have some assurance that we shall meet a population described in terms of such averages in definable circumstances. In fact, we are rarely, if ever, able to do so in branches of enquiry relying on this procedure.

(3) Laboratory or observatory experience may endorse with some plausibility the assumption that successive uncontrollable instrumental errors turn up randomwise; but the assumption that deviations of population scores from a fictitious mean are random variables derives its sanction wholly from the arbitrary and Platonic concept of the infinite hypothetical population. This concept has not gained universal acceptance, being indeed rejected by the school most antagonistic to a behaviourist approach to the theory of probability.

(4) Though the theory of regression makes use of the same

formal algebra as the Gaussian theory of error, it derives no sanction from whatever claims to usefulness we concede to the latter. Its aim is different. The concept of law which it embraces is different. Its factual assumptions are different.

*The Grammar of Science.* If we enlist in the undertaking to which the calculus of exploration commits us, all we can thus hope to gain at the end of the journey is a relation between averages referable to populations beyond our powers to recreate. In the domain of social relations, the methods it prescribes can never give us the assurance of unmasking a causal nexus; and the recognition of such a causal nexus is a necessary basis for applying our knowledge constructively. Under a pretentious edifice of mathematical sophistication, we enter the hope of changing our world. We resign ourselves to the role of the passive and helpless spectator. Having buried the Baconian formula, we must then doctor our definition of scientific law accordingly, and regression takes its place in a formalised ideology. What might otherwise seem to be misconceptions too trivial to merit rebuttal are indeed by-products of such an ideology set forth in Pearson's *Grammar of Science*.

As one of the last survivors of a generation which received it as a new evangel, I find it hard to discover any rational ingredients in an enthusiasm which I once shared with so many others. As one sees it at a more mature age, Pearson's message is consistent with what the rising generation now interpret as its main thesis. Briefly, this is:

(a) the raw data of science are individual mental images in a static framework which excludes the recognition of unique historical events as a proper theme for intelligent reflection;

(b) the concept of scientific law is reducible to one "brief statement or formula which resumes the relationship between a group of facts."

On a generation whose Victorian elders welcomed the late Professor Henry Drummond's *Natural Law in the Spiritual World* as a challenging contribution to human thought, the brevity of the aforesaid statement exerted an appeal a younger generation can never recapture. In these days we hear enough, and it

may be too much, about *operational* research. So we do not commonly expect to specify a particular assemblage of facts by a relationship which is unique in the sense that it is independent of considerations relevant to the chosen framework of classification. How then are we to specify the terms of reference of the relationship expressed by the brief statement or formula? The student of today will no longer be content with the answer that the formula must be short. He or she will ask: what can it do for us? If the answer we seek is Pearson's own view, we may well infer that the composite photograph of the deserving recipient is Quetelet's normal man at home by his normal fireside with his normal wife and normal offspring at the end of a normal day; but we shall certainly not make his acquaintance in a world outside and antecedent to ourselves. We shall locate him (p. 110, *vide infra*) in the Platonic domain of our individual *perceptions*.

Even the definition of *brief* in this context is vague. Pearson nowhere explicitly asserts that the terms of reference of his brief formula are necessarily numerical; but he is very clearly of one mind with Kelvin when the latter declares:

When you can measure what you are talking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind, it may be the beginning of knowledge but you have scarcely in your thoughts advanced to the stage of science whatever the matter may be.

The concluding remarks of his discussion of the content of natural, in contradistinction to civil or spiritual, law affirm an out-and-out idealistic viewpoint *en rapport* with his adherence to the principle of insufficient reason, as quoted elsewhere. He there defines the *Weltanschauung* of Pearsonian man in the following terms (p. 110):

He recognises that the so-called law of nature is but a simple *résumé*, a brief description of a wide range of his own perceptions and that the harmony between his perceptive and reasoning faculties is not incapable of being traced to its origin. Natural law appears



to him an intellectual product of man, and not a routine inherent in "dead matter." The progress of science is thus reduced to a more and more complete analysis of the perceptive faculty—an analysis which unconsciously and not unnaturally, if illogically, we too often treat as an analysis of something beyond sense-impression. Thus both the material and the laws of science are inherent in ourselves rather than in an outside world.

A view of science so conceived can readily accommodate such artefacts as the normal man and the normal environment cheek by jowl with Ricardo's economic man, the eternal verities of Malthus and such scholastic generalisations as the so-called law of supply and demand. Thus it extends the benefits of free grace on Kelvin's terms to the newer disciplines, which the great grammarian woos with temporary disregard for the foregoing renunciation of the external world, as when he declares (p. 12):

The unity of all science consists alone in its method, not in its material. The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequence, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the method in which they are dealt with. The material of science is co-extensive with the whole physical universe, not only that universe as it now exists, but with its past history and the past history of all life therein.

In such temporary lapses into a conciliatory attitude towards the claims of non-quantitative historical studies, Pearson does not disclose how the chronologically closed field of our perceptions is able to accommodate what antedated their existence; and indeed part of his appeal resides in an eclecticism which entitles the reader of any persuasion to find in the *Grammar*, as in Engels' *Anti-Duehring*, something to his or her taste. At a time when the 1902 Education Act had rekindled the embers of the Religious Tests Controversy, he is able to allay our suspicion that he is defending a metaphysic in the tradition of so staunch a churchman as Berkeley by a disarm-

ing bouquet to the bishop's opponents in the same context (p. 109):

It may seem to the reader that we have been discussing at unjustifiable length the nature of scientific law. Yet therein we have reached a point of primary importance, a point over which the battles of systems and creeds have been long and bitter. Here the materialists have thrown down the gauntlet to the natural theologians, and the latter in their turn have endeavoured to deck dogma with the mantle of science.

No doubt many other circumstances were propitious to the uncritical welcome the book received from the Edwardian left and right. One is certainly the fact that the author had a magnificent command of pungent English; and his early volume of essays entitled *The Ethic of Free Thought* is still readable as a model of prose style. Above all, a generation of biologists weary of reckless and fruitless speculations about phylogeny in the wake of the Darwinian controversy and goaded by Kelvin's challenge in the setting of spectacular inventions traceable to discoveries of physicists, was eager enough to welcome any prospect of exalting the claims of its own field of enquiry to rank as *exact* science. Pearson's repudiation of the Baconian viewpoint extended to biology and to sociology the benefits of mathematical free grace at his own price. What price psychology has paid for the privilege will be the theme of the next chapter.

## THE IMPASSE OF FACTOR ANALYSIS

A RETROSPECTIVE GLANCE at the pages of *The Grammar of Science* is a fitting introduction to factor analysis, as will appear from the following profession of faith by Thurstone (*The Vectors of the Mind*) :

A scientific law is not to be thought of as having an independent existence which some scientist is fortunate to stumble on. A scientific law is not a part of nature. It is only a way of comprehending nature . . . the chief object of science is to minimise mental effort. . . . It is in the nature of science that no scientific law can ever be proved to be right. It can only be shown to be plausible. The laws of science are not immutable. They are only human efforts towards *parsimony in the comprehension of nature.* (*Italics inserted.*)

Here we have the Pearsonian evangel in its entirety. Of two or more descriptions of natural phenomena, we shall deem to be acceptable the one which is most *economical*. It will thus be less difficult for the reader to discern how largely the Pearsonian formula for natural law derives its urgency from the inclination to equip the Pearsonian calculus of exploration with a plausible rationale. By a familiar feat of metalepsis congenial to the professional mathematician, the most parsimonious form of statement is that which embraces a *minimal* solution. The method of least squares will lead us to the true description of concomitant variation of tuberculosis rates and of housing density, because it *minimalises* the variance of the deviations around a line fitted to a scatter diagram. Accordingly, we abandon the obligation to ask what useful outcome, if any, is the reward of our labours or to subject the conclusions to which the method leads us to the arbitrament of verification by recourse to other procedures.

It is one of time's ironies that the statistical technique most heavily indebted to the essentially novel feature of Pearson's contributions to statistical theory, namely the *product-moment*

index of concomitant variation, has led after forty years of fruitless wandering in the wilderness of matrix algebra to a *cul de sac* in which its practitioners are at length asking: is there indeed any unique sense in which a statement summarising our acquaintance with natural phenomena is as economical as may be? Meanwhile, the influence of Pearson has extended over a far wider field. We necessarily think of algebraic formulae as more economical than verbal statements because in one sense they are so. Accordingly, observations upholstered with sufficient algebraic sophistication are *ipso facto* praiseworthy on that account. By the same token, we discount the contribution of the naturalist or of the physician with clinical judgment, unless the presentation of the observational record conforms to a ritual whose credentials lie outside the curriculum of his training. More and more, he thus learns to rely on interpretation by recourse to methods which he understands less and less.

Thus the influence of Pearson has brought about a new orientation of values. When the *Grammar of Science* came out in its first edition, T. H. Huxley and his influential following could still boast that the man of science takes nothing on trust. To the naturalist smarting under clerical antagonism to assertion of the right to indulge in the luxury of speculation unlicensed by scriptural sanction and contrary to the customary interpretation of the *ipsissima verba* of Holy Writ, there was nothing incongruous in this claim; but no reflective person of our own generation could regard it as now true. Indeed, few of the younger men of science would endorse it as the statement of a congenial ideal. A vast literature on *Factor Analysis* accumulated in the last forty years bears witness to the role of Pearson as the parent of a methodology fitting to the authoritarian temper of our time. In every, or almost every, teachers' training college throughout the English-speaking world, students whose mathematical equipment rarely transgresses the boundaries of the high school curriculum listen to glib parables about the rotation of axes in hyperspace with docile anticipation of examinations conducted in a temper as likely to encourage a rational scepticism as did the forced repetition of the lesser catechism to Scottish children of an earlier generation.

The impact of this peculiar private language of so-called *factor-space* on the uninitiated recalls the bewilderment of schoolboys of an earlier generation when their instructors failed to distinguish between the ether as a convenient fiction to accommodate the algebra appropriate to the description of actual wave motion and the ether as a self-contradictory metaphor for the actual medium in which propagation amenable to description in such terms takes place. Nearly all current accounts of factor analysis presuppose some elementary acquaintance with matrices and determinants or start with an exposition of grid algebra, leaving the initiate with the impression that the credentials of the procedure are incomprehensible in any other idiom. The truth is that every logical issue raised by the claims of factor analysis is understandable, and the more readily so, if we keep a firm foothold in classical models for which high school algebra suffices. The need for an alternative notation arises only when the work of computation otherwise becomes excessively laborious, and then only for that reason.

In this chapter, we shall trace to its origin the *impasse* into which factor analysis has led contemporary psychologists who have followed this trail. It will not be necessary to presume a knowledge of matrix algebra, but it will be necessary at the outset to introduce a few terms to avoid periphrasis in what follows. By the *correlation coefficient* we shall consistently signify the Pearsonian covariance (*product-moment*) formula, denoted  $r_{ab}$  for paired measurements of attributes A and B (e.g. height of son and height of father or the A-test score of an individual and the B-test score of the same individual). In general terms, Factor Analysis is the attempt to interpret observations which disclose an appreciable measure of correlation, in the sense that  $r_{ab} \neq 0$ , when our data embrace several pairs of such attributes.

From variable measurements referable to attributes A, B, C, D . . . , etc., we may calculate correlation (product-moment) coefficients  $r_{ab}, r_{ac}, r_{ad}, r_{bc}, r_{bd}, r_{cd} \dots$  If we lay these out grid-wise in accordance with the pattern shown overleaf, it is customary to speak of the symmetrical arrangement as the *correlation matrix* for the set of attributes.

By a suitable choice of what we choose to label as A, B, etc., it may then be possible to lay out such a symmetrically ordered grid with successive cell entries of each column (from top to bottom) and successive cell entries of each row (from left to right) in ascending numerical order. If this is indeed possible, we say that the correlated attributes conform to the *hierarchical*

	A	B	C	D	...
A	...	$r_{ba}$	$r_{ca}$	$r_{da}$	
B	$r_{ab}$	...	$r_{cb}$	$r_{db}$	
C	$r_{ac}$	$r_{bc}$	...	$r_{dc}$	
D	$r_{ad}$	$r_{bd}$	$r_{cd}$	...	
...	...	...	...	...	...

principle. From data w.r.t. paired scores of groups of individuals each subjected to a set of educational tests A, B, C, etc., Spearman (1912) claimed that it was possible to exhibit such a hierarchical pattern; and advanced the view that the inter-correlations then arise from concomitant variation assignable to a single source or factor, which he named *g* and identified as *general intelligence*.

Of itself, the fact that  $r_{ab}$ ,  $r_{ac}$  and  $r_{bc}$  appreciably exceed zero, as is implicit in the statement that A, B, C are intercorrelated variables, merely signifies that the scores  $x_a$ ,  $x_b$ ,  $x_c$  of one and the same individual for tests A, B, C agree too closely to be consistent with randomwise distribution. We may regard this as indicative of the existence of some source of individual variation common to any pair of tests; but this is consistent with the possibility that there are as many common sources (*factors*) as the number of different pairs, i.e.  ${}^nC_2$  for  $n$  tests. Factor Analysis starts from the assumption that the correct interpretation of the properties of the correlation matrix is that which is consistent with postulating the *minimum* number of factors common to one or more pairs. Accordingly, it seeks an answer to two questions: (a) what is the minimum number of factors we require on that understanding; (b) how can we assess the contributions of each of them to variation w.r.t. the individual's score for each test?

If we can say that *one* common factor, e.g. Spearman's *g*,

provides the answer to the first question last stated, we can answer the second decisively in terms consistent with the arbitrary choice of an appropriate stochastic model. Otherwise, no stochastic model can certainly lead us to the goal we seek. There is indeed more than one current prescription for *multiple* factor analysis, and different prescriptions lead us to assign different values to the contribution of the several factors of the minimum set. Thus Spearman's original formulation, if applicable in practice and acceptable in theory, confers no licence on the subsequent development of the statistical theory he propounded.

When the factor pattern is as described above, the stochastic model which fulfils the expectations implicit in Spearman's interpretation of the correlation matrix for a set of inter-correlated test scores with a single common factor is a more general case of the *umpire bonus model* than the one discussed in our last chapter. There each player receives the same bonus from the umpire. As before, we shall postulate that each player tosses a die and that the umpire does likewise at each game; but we shall suppose that the bonus each player incorporates with his individual score ( $x_{a.0}$ ,  $x_{b.0}$ , etc.) in his final score ( $x_a$ ,  $x_b$ , etc.) is some multiple of the umpire's score, such that the actual bonus any one player receives in one and the same game is different from that which any other receives.

For a set-up involving one umpire and four players (A, B, C, D) the definitive equations of the score are thus

$$x_a = A \cdot x_u + x_{a.0}$$

$$x_b = B \cdot x_u + x_{b.0}$$

$$x_c = C \cdot x_u + x_{c.0}$$

$$x_d = D \cdot x_u + x_{d.0}$$

If  $V_u$  and  $V_a$  respectively denote the variance of the umpire score distribution and that of the final score of player A, the properties of the model relevant to our theme are, as derived in *Appendix II*:

$$r_{ab} = r_{au} \cdot r_{bu} ; r_{au} = A \cdot V_u^{\frac{1}{2}} \cdot V_a^{-\frac{1}{2}} ; V_a = A^2 V_u + V_{a.0} \quad (i)$$

Thus the fraction (so-called *communality*) of the variance of the

distribution of the final score of player A contributed by the umpire bonus is:

$$\frac{A^2V_u}{V_a} = r_{au}^2 \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

In virtue of the relation  $r_{ab} = r_{au} \cdot r_{bu}$ , we may write

$$r_{ab} \cdot r_{cd} = r_{au} \cdot r_{bu} \cdot r_{cu} \cdot r_{du}$$

$$r_{ac} \cdot r_{bd} = r_{au} \cdot r_{cu} \cdot r_{bu} \cdot r_{du}$$

$$r_{ad} \cdot r_{bc} = r_{au} \cdot r_{du} \cdot r_{bu} \cdot r_{cu}$$

$$\therefore r_{ab} \cdot r_{cd} = r_{ac} \cdot r_{bd} = r_{ad} \cdot r_{bc} \quad . \quad . \quad . \quad (iii)$$

If  $r_{au} < r_{bu} < r_{cu} < r_{du}$ , it follows that

$$r_{ab} < r_{ac} < r_{ad}$$

$$r_{ab} < r_{bc} < r_{bd}$$

$$r_{ac} < r_{bc} < r_{cd}$$

$$r_{ad} < r_{bd} < r_{cd}$$

This is Spearman's *hierarchical principle*.

We may now recognise the following rule:  $r_{au}^2$  is the ratio of the sum of all the products of pairs of observed correlations involving the A-score to the sum of all the correlations which do not involve the A-score. Thus

$$r_{ab} \cdot r_{ac} + r_{ab} \cdot r_{ad} + r_{ac} \cdot r_{ad} = r_{au}^2 (r_{bu} \cdot r_{cu} + r_{bu} \cdot r_{du} + r_{cu} \cdot r_{du})$$

$$r_{bc} + r_{bd} + r_{cd} = r_{bu} \cdot r_{cu} + r_{bu} \cdot r_{du} + r_{cu} \cdot r_{du}$$

From this relation we can get a pooled estimate of  $r_{au}$  and in the same way of  $r_{bu}$ ,  $r_{cu}$ , etc. From these estimates we may then reconstruct a correlation matrix of *estimated* values of  $r_{ab}$ ,  $r_{ac}$ , etc., as shown opposite.

The identity embodied in (iii) embraces a more general rule when we have before us the scores of more than 3 players, as in the foregoing exemplary correlation matrix. For any pair of columns we can pick out two rows without zero entries, e.g.:

$$\begin{array}{cc} r_{ac} & r_{bc} \\ r_{ad} & r_{bd} \end{array} ; \quad \begin{array}{cc} r_{ab} & r_{cb} \\ r_{ad} & r_{cd} \end{array} ; \quad \begin{array}{cc} r_{ab} & r_{bd} \\ r_{ac} & r_{cd} \end{array} ; \quad \begin{array}{cc} r_{ab} & r_{ac} \\ r_{bd} & r_{cd} \end{array} \dots\dots$$



# THE IMPASSE OF FACTOR ANALYSIS

For each of these *tetrads* (Spearman), spoken of as a *determinant minor of order 2* in the idiom of matrix algebra, the difference

	$r_{au}$	$r_{bu}$	$r_{cu}$	$r_{du}$
$r_{au}$	...	$r_{au} \cdot r_{bu}$ $= r_{ab}$	$r_{au} \cdot r_{cu}$ $= r_{ac}$	$r_{au} \cdot r_{du}$ $= r_{ad}$
$r_{bu}$	$r_{au} \cdot r_{bu}$ $= r_{ab}$	...	$r_{bu} \cdot r_{cu}$ $= r_{bc}$	$r_{bu} \cdot r_{du}$ $= r_{bd}$
$r_{cu}$	$r_{au} \cdot r_{cu}$ $= r_{ac}$	$r_{bu} \cdot r_{cu}$ $= r_{bc}$	...	$r_{cu} \cdot r_{du}$ $= r_{cd}$
$r_{du}$	$r_{au} \cdot r_{du}$ $= r_{ad}$	$r_{bu} \cdot r_{du}$ $= r_{bd}$	$r_{cu} \cdot r_{du}$ $= r_{cd}$	...

between the cross products is zero in accordance with (iii) above. For instance,

$$r_{ac} \cdot r_{bd} = r_{au} \cdot r_{bu} \cdot r_{cu} \cdot r_{du} = r_{bc} \cdot r_{ad}$$

$$\therefore r_{ac} \cdot r_{bd} - r_{bc} \cdot r_{ad} = 0 \quad . \quad . \quad . \quad (iv)$$

In the language of matrix algebra, we express this formulation of the hierarchical principle by saying that *every second order minor of the correlation matrix vanishes*.

In so far as  $r_{au}$ ,  $r_{bu}$ , etc., are adequate summarising indices the cell entries of the foregoing grid should tally closely with the observed values of  $r_{ab}$ ,  $r_{ac}$ , etc. To get the relevance of our model to the use of factor analysis into focus, we shall now suppose that:

(a) we are able to observe a very large number of games in which the 4 players A-D participate;

(b) we are able to record the final score of each player on each occasion, but know neither what bonus the umpire contributes to each nor what are the individual scores of the players;

(c) we regard the final scores of any two players at one

and the same game as a pair for the purpose of determining score correlations  $r_{ab}$ ,  $r_{bc}$ , etc.;

(d) the calculation yields a set of values consistent with the hierarchical principle, and the values of  $r_{ab} \cdot r_{cd}$ ,  $r_{ac} \cdot r_{bd}$  and  $r_{ad} \cdot r_{bc}$  are approximately equal;

(e) the cell entries of the correlation matrix reconstructed from our estimates of  $r_{au}$ , etc., tally closely with the observed values of  $r_{ab}$ , etc.

(f) our estimated communalities conform to the order

$$r_{au} < r_{bu} < r_{cu} < r_{du}.$$

In this situation, the fact that the scores of any two players are correlated indicates that each receives a bonus which varies from game to game, but not necessarily randomwise. That the score tetrads are approximately equal, and that the hierarchical principle holds good, point to the existence of a single source of the bonus each player receives and to its randomwise distribution from game to game. What then may we also infer from the border factors of our grid? From (ii) we see that  $r_{au}^2$  represents the proportionate contribution of the umpire's bonus to the variance of the score distribution of player A. The statement that  $r_{au}$  is the smallest of the factors means that the withdrawal of the umpire's bonus would proportionately diminish the dispersion of the score of A *less* than that of the other players. Conversely, the statement that  $r_{du}$  is the greatest of the factors signifies that the withdrawal of the umpire's bonus would proportionately diminish the dispersion of the score of D more than that of any of the other players.

Let us now assume that our model reproduces the essential features of the application of 4 tests A-D to a large number of boys and girls. Each boy or girl then corresponds to a game. The test score A and the test score D of one and the same boy or girl corresponds to a pair of players' final scores in a particular game of the sequence. We deem the umpire bonus to represent the common ability which tests A and D assess, and the players' individual scores to represent the specific abilities which the tests respectively assess and/or *error*. The term error in this context is equivocal, meaning the unreliability of the

test in terms of: (a) mistakes the record may contain; (b) inconsistencies w.r.t. individual response to one and the same stimulus on different occasions.

Thus  $r_{au}^2$  signifies the proportionate contribution of the common ability to the dispersion of the A test score in the population tested. If  $r_{au}$  is less than  $r_{bu}$ ,  $r_{cu}$ ,  $r_{du}$ , this signifies that the common ability assessed by the 4 tests contributes proportionately least to the variation of the A test score in the population as a whole. By the same token we regard the A-test as a less sensitive measure of individual variability w.r.t. the assumed common ability.

If we concede all the assumptions provisionally adopted, we may thus state the ostensible uses of the procedure outlined. It claims to tell us:

- (i) whether the correlation between three or more measurements is referable to a single common component;
- (ii) whether one test which putatively assesses such a single common component is more or less sensitive in the sense that it yields results less influenced by other sources of variation.

The second of the foregoing is of theoretical interest only as a more or less useful means of sharpening the tools of the research worker. What precisely we may achieve by the first is an empirical issue, since the procedure under discussion does not guarantee to label the common component usefully. As stated, it led Spearman to conclude that there is a single component ability (*g*) contributory to the individual's rating referable to different types of scholastic tests deemed to assess what we commonly call intelligence. We may interpret this conclusion, if true, by saying that there is some meaningful nucleus in the use of the adjective *intelligent*. If so, we may deem it desirable to make a linguistic recommendation in favour of restricting the use of the word *intelligent* to the individual's rating on a so-called *g*-saturated test, i.e. a test *T* such that the *g*-communality  $r_{tg}^2$  accounts for a very high percentage of the test-score variance.

What is not clear is whether we have accomplished more than refinement of our terminology by doing so. That such

refinement may be eminently useful at a certain stage in the development of any science, especially a young science, is not disputable; but the usefulness of a procedure to promote refinement of a taxonomy does not suffice to justify the claim that it automatically generates a correct hypothesis without demanding any initial statement of the job assigned to it. What we shall deem to be useful in an operational sense on all fours with the criteria we adopt in experimental physiology will be what leads us to the recognition of regularities of human behaviour verifiable by other means. Only if susceptible to independent verification, can the conclusions to which our selected stochastic model leads us, justify its endorsement, as when we justifiably accept the postulate of randomwise association of the gametes to interpret hereditary transmission.

Be this as it may, we may concede that the stochastic interpretation of a hierarchical factor pattern is an attractive one, and is not one we should lightly dismiss as a clue to follow up, if such patterns turned up commonly in the course of statistical investigations on intra-group variation. The truth is that subsequent research on the lines first explored by Spearman has led to the conclusion that a recognisably and consistently hierarchical pattern rarely, if ever, emerges from comparison of scores referable to different tests or to measurements referable to different attributes of different individuals or—Cattell's *P*-technique (p. 229)—to successive states of the same individual. While some workers in the same field have recorded plausible examples of bi-factor patterns, i.e. hierarchical systems of inter-correlations interpretable in terms of one factor common to all test scores and other factors common to discrete groups of test scores in a battery, the procedures subsumed under the term *multiple factor analysis* derive no such plausibility from immediate inspection of the data. As stated, their aim is to assign *loadings*, such as  $r_{au}^2$ ,  $r_{bu}^2$  in the foregoing, to components of test score variance without invoking the assumption that any such loading exceeds zero for all classes of scores involved.

Before we ask how far any unique stochastic model can endow this pursuit with some similitude of cogency, let us elaborate the foregoing model situation. We shall now suppose that each of more than two umpires ( $U_1$ ,  $U_2$ , etc.) contributes

to some multiple of each player's individual score some multiple of his own ( $x_1, x_2$ , etc.); but we shall not exclude the possibility that any multiplicative constant ( $F_1, F_2$ , etc.) may be zero. The player's individual score on the  $F$ th test will therefore be

$$x_f = F_1 \cdot x_1 + F_2 \cdot x_2 \dots + F_0 \cdot x_{f0} \quad . \quad (v)$$

By the same reasoning (*Appendix II*) employed in deriving the formal relations implicit in the rule of the game when there is only one umpire bonus, we then derive for the set-up involving two umpires the following relations involving the observable correlation ( $r_{fg}$ ) between the player's score on two tests (F, G) and the correlation between either test score of the player and that of either umpire:

$$r_{f1} \cdot r_{g1} + r_{f2} \cdot r_{g2} = r_{fg}$$

$$F_1^2 \cdot V_1 + F_2^2 \cdot V_2 + F_0^2 \cdot V_{f0} = V_f$$

$$r_{f1}^2 + r_{f2}^2 + \frac{F_0^2 \cdot V_{f0}}{V_f} = 1$$

More generally for  $n$  umpires we may write:

$$r_{fg} = \sum_{x=1}^{x=n} r_{fx} \cdot r_{gx} \text{ and } V_f = F_0^2 \cdot V_{f0} + \sum_{x=1}^{x=n} F_x^2 \cdot V_x \quad . \quad (vi)$$

By recourse to the expression on the left hand in (vi) we may derive without much labour a formula analogous to (iv) for the correlation matrix when only two umpires participate, viz.:

$$\begin{aligned} r_{ad} (r_{bc} \cdot r_{cf} - r_{bf} \cdot r_{ce}) - r_{ac} (r_{bd} \cdot r_{cf} - r_{bf} \cdot r_{cd}) \\ + r_{af} (r_{bd} \cdot r_{ce} - r_{cd} \cdot r_{be}) = 0 \quad . \quad (vii) \end{aligned}$$

The last equation cited does not obviously conform to a pattern of which (iv) is a special case, unless the reader is familiar with the notation of matrix algebra. In the idiom of matrix algebra (*Appendix III*), it signifies that all determinant minors of order 3 in the correlation matrix vanish for 2 umpires as all determinant minors of order 2 vanish for the foregoing case of one umpire; and more generally we may derive from

(v) by recourse to determinants the rule that all minors of order  $(n + 1)$  vanish when there are  $n$  umpires.

With the enunciation of this rule by Thurstone, theory loses all foothold in the comparatively firm soil of the classical model, and we thread our way through a maze of algebraic metaphors invoked to describe subsequent symbolic manipulations rather than the properties of the model itself. Since but a small fraction of research workers who rely on multiple factor analysis as a so-called tool of research are at home in the hyperspace of the matrix notation, the ensuing confusion of thought is what one might well anticipate. Albeit the labour involved in discussing more than two common factors by recourse to schoolbook algebra is a sufficient justification for preferring the use of matrices, the invocation of three or more factors involves no logically new issue which the unrestricted invocation of two wholly or partially common factors in place of Spearman's single one cannot bring clearly into focus.

Accordingly, we shall now relinquish the privilege of taking a backstage view of the game. All we shall permit ourselves to know is the final test score of each player; and our task in the Spearman tradition will be both to decide how many umpires participate and what contribution the variance of the score distribution of each umpire makes to that of the distribution of the final score of each player. Even if we then have the assurance that the rules of the game are correctly subsumed in (v) above, we are in a quandary. The principle last stated may satisfy our requirements for the first part of our task; but we then find that the evaluation of the loadings ( $r_{f1}^2$ ,  $r_{f2}^2$ , etc.) admits of no unique solution, unless we rely on assumptions deriving no sanction from the observations themselves.

This will be sufficiently clear, if we now divert our attention from the putatively appropriate model to the test situation. We then have to take cognisance of our initial arbitrary assumption that the score components are strictly additive as well as the equally arbitrary assumption that their distribution is random-wise. It is difficult to see why the possibility that all minors of order  $n$  in a correlation matrix vanish in theory and have negligible numerical values in practice should endorse either the one or the other; but if we grant the assumption that we

really know the rules of the game in general terms as stated above, we come face to face with difficulties more disturbing to the docile practitioner. It will suffice to mention one of them.

In Spearman's original formulation there are two factors only, one common to all test scores, the other unique and putatively referable to a specific attribute measured by a given test after correction of the test scores for error in Spearman's sense, i.e. reliability of the test procedure. Now the foregoing formulation places *no* restriction on the value of the constant  $F_0$ . Thus we are free to reject the proposition that any test score has a truly *specific* component; and it is here that prescriptions for evaluating the loadings respectively given by Thurstone's following and that of Hotelling diverge radically. Thurstone endows every test score with, and Hotelling deprives every test of, a specific component.

Hotelling's position is wide open to criticism for two reasons. Of minor importance is the fact that any test must be subject to independent variability unless perfectly reliable, and we have therefore to assume without opportunity for verification that correction for unreliability, being itself a statistical device and therefore fallible on its own terms, removes all sources of error in the widest sense of the term. From the viewpoint of the pure mathematician this has the merit of empowering us to extract a tidy solution implicit in the right-hand expression of (v) above. Since all the variance of the test score is referable to the communalities in the absence of specifics, the variance of the score distribution for each test is exhaustively expressible in terms of them. To guarantee a unique evaluation of the communalities, we then make the further assumption that the total number of wholly or partially common factors assessed by  $r_{f1}^2$ ,  $r_{2f}^2$ , etc., is equal to the number of tests. The addition of another test to the battery must then change, and may grossly change, our previous evaluation of each test as a gauge for the abilities referable to the factors.

Godfrey Thomson (*Factorial Analysis of Human Ability*, p. 68) refers to this quandary in the following remarks (*italics inserted*):

. . . If there are, say twenty tests, there will be twenty principal axes ranging from longest to shortest, and twenty Hotelling compo-

nents. But the first four or five of these will go a long way towards defining a man's position in the tests, and will do so better than any other equally numerous set of factors, whether of Hotelling's or of any other system. In this respect Hotelling's factors undoubtedly stand foremost. They will not, however, reproduce the *correlations* exactly unless they are all used, whereas in Thurstone's system a few common factors can, theoretically, do this, though in actual practice the difference of the two systems in this respect is not great. *The chief disadvantage of Hotelling's components is that they change when a new test is added to the battery.*

The gentleness of the rebuke in the last sentence cited will reinforce its cogency, if read in a reflective temper. One might with equal propriety say that a method of analysis is admissible if it yields consistent figures for percentage composition in repeated titrations of  $x$  cc. of a solution, but with the trifling drawback of yielding a totally different estimate based on repeated titrations of  $y$  cc. of one and the same solution. We need therefore record no despondency if another school of factor theory repudiates both assumptions on which Hotelling relies.

From the foregoing it should be clear that Hotelling's formulation does not embrace Spearman's as a special case. That of Thurstone admits no test  $F$  for which  $F_0 = 0$  and embraces Spearman's hierarchical pattern when the value of  $F_x$ , etc., for all values of  $x$  from 1 to  $n$  is zero for every test score component other than the particular ( $r$ th) common component to which  $F_r$  is referable; but the endorsement of a component specific to each test does not lead to a unique evaluation of the loadings, nor does it clearly guarantee that addition of a new test to the battery leaves our previous evaluation of loadings intact. Again I quote from Godfrey Thomson (*op. cit.*):

Whether Thurstone's common factors will remain invariant in augmented batteries, and if so under what conditions, is a question we shall consider at a later stage in this book. Though such invariance seems unlikely, it is not obviously inconceivable. . . . The Hotelling components . . . can be calculated exactly from a man's scores, whereas Spearman or Thurstone factors can only be estimated. This is because the Hotelling components are never more numerous than the tests, whereas the Thurstone or Spearman



factors, including the specifics, are always more numerous than the tests. For the Hotelling components, therefore, we always have just the same number of equations as unknowns, whereas we have more unknowns than equations in the Spearman-Thurstone system.

To formulate his own rule for the evaluation of the loadings Thurstone relies on a somewhat personal interpretation of the familiar dictum of William of Occam, whence the special relevance of the citation at the beginning of this chapter. One way of stating the Thurstone principle of economy is as defined by Godfrey Thomson, viz. that the most economical evaluation maximises the contribution of the specifics and minimises the number of contributory common factors in the balance sheet of the test score variance. Burt expresses Thurstone's approach in different terms (*The Factors of the Mind*, p. 162):

His interpretation of the "law of parsimony in scientific description" requires, not (as mine does) that each factor in turn should account for the greatest possible amount of the variance, but that: (a) the total number of factors entering into the whole set of traits and (b) the number of factors entering into each sample trait should be as small as possible. . . . He therefore seeks a factorial matrix in which every factor shall have at least one zero coefficient for at least one of the tests.

The controversy between the different schools of factor theory as revealed in the passage last quoted thus lays bare an issue raised elsewhere and at the beginning of this chapter. The *Grammar* of Karl Pearson propounded an ideology tailored to the requirements of his conviction that the edifice he erected on the foundations laid by Quetelet could accommodate all the future requirements of truly scientific enquiry pertaining to evolution, man's nature and human society. The keystone of the edifice was the author's own interpretation of *entia non multiplicanda praeter necessitatem*, a counsel of modest prudence and salutary enough unless beatified as an article of faith. The truth is that there are as many criteria of economically interpreting nature as there are different frameworks of interpretation dictated by different ways in which we may seek to bind nature in the service of man. The present dilemma of factor analysis arises from the fact that its initial terms of reference

embrace a classificatory task without explicit formulation of the end in view, an explicit admission of protagonists\* anxious to promote its claims to consideration in the domain of experimental science.

That the real dilemma arises from the impossibility of uniquely defining a universal criterion of economy or parsimony in the interpretation of nature is at least evident to one contemporary, who has devoted many years to evaluating the credentials of factor analysis, as when Godfrey Thomson (*op. cit.*, pp. 133-34) writes:

*Shorthand descriptions.* It is to be observed that an analysis using the minimal number of common factors, and with maximised specific variance, is capable of reproducing the correlation coefficients exactly by means of these few common factors, and in the case of an artificial example will actually do so; while in the case of an experimental example including errors, it will do so at least as well as any other method. If this is our sole purpose, therefore, the Thurstone type of analysis is best, since it uses fewest factors.

But the few common factors of a Thurstone analysis do not enable us to reproduce the original test scores from which we began, they do not enable us to describe all the powers of our population of persons very well. With the same number of Hotelling's "principal components" as Thurstone has of common factors we could arrive at a better description of the scores, though a worse one of the correlations. The reader may reply that he does not want factors for the purpose of reproducing either the original scores or the original correlations, for he possesses these already! But what we really mean, and what it is very convenient to have, is a concise shorthand description, and the system we prefer will depend largely on our motives, whether we have a practical end in view or are urged by theoretical curiosity. The chief practical incentive is the hope that factors will somehow enable better vocational and educational predictions to be made. Mathematically, however, as we have seen, this is impossible. If the use of factors turns out to improve vocational advance it will be for some other reason than a mathematical one. For vocational or educational prediction means, mathematically, projecting a point given by  $n$  oblique co-ordinate axes called tests on to a vector representing the occupation, whose angles with the tests are known, but which is not in the  $n$ -space of the tests. The use

\* See Cattell and Williams, p. 229.

of factors merely means referring the point in question to a new set of co-ordinate axes called factors, a procedure which cannot define the point any better, and, unless care is taken, may define it worse, nor does the change of axes in any way facilitate the projection on to the occupation vector.

Against the background of these remarks it is suggestive to contrast the role of economy with the role of analogy in the search for truth. Most of us will agree that reasoning from analogy in matters pertaining to human affairs often leads to gross error and rarely, if ever, to a profitable outcome. None the less, the progress of experimental physics during the four centuries which have followed the exposition of Gilbert's *terella* is in large measure due to the exploitation of daring metaphors whose initial plausibility we cannot easily discern at a distance. Thus we may well ask why analogy which is so good a servant of natural philosophy is so bad a master of humanistic enquiry. If we approach the issue from what G. P. Meredith calls the epistemic viewpoint the question admits of more than one answer, but an answer relevant to our present theme is not far to seek. The physicist follows an analogy only in so far as it leads him to conclusions susceptible to factual verification. Where two analogies suggest different conclusions, he designs the *experimentum crucis* to adjudicate on their respective merits. I have yet to discover that comparisons between the State and the organism or between the nervous system and the self-guided missile have led to conclusions sufficiently free from ambiguity to succumb to disproof, if false.

To me it seems that the peculiar status of the principle of parsimony in Pearson's system is on all fours with the misuse of analogy. Few of us, if any, will prefer a more laborious interpretation of a physical event to a more simple one, when both conform to factual requirements; but if such was indeed the motivation of those who first explored the consequences of the heliocentric interpretation of planetary motion, it would be false to say that the world of science rejected the Ptolemaic view for this reason only. In his own time, Hipparchus rejected the heliocentric viewpoint endorsed by Aristarchus, because no annual parallax of a star was detectable by methods then available. The doubt remained

(p. 187) till the thirties of the last century, but the view of Aristarchus meanwhile gained strength from a succession of other discoveries, made after Copernicus revived it, viz. the existence of Jupiter's moons, the retardation of the pendulum by latitude, the flattening of the earth at the poles and the phenomenon of stellar aberration.

Thus it is not true to declare that astronomers rejected the Ptolemaic system by reliance on any rule of thumb application of the renowned razor of Occam. Were it otherwise, they would have done so with some misgivings. For the decision to do so raises a dilemma of the sort disclosed in the foregoing citation from Godfrey Thomson. Admittedly, the heliocentric postulates lead to a simpler method for tracking the course of the planets, a consideration of some practical interest in the milieu of Copernicus, when navigators had to rely on planetary conjunctions and occultations for determining the longitude of a ship at sea. Contrariwise, the geocentric view leads to a simpler method of representing the position of the fixed stars on which the mariner still relies to fix his latitude. Accordingly, all current books on nautical astronomy still adhere to the Ptolemaic conception of the celestial sphere, except in the chapter or chapters devoted to planetary motion.

In so far as a proponent of multiple factor analysis may, as does Thurstone, invoke the history of astronomy to endorse the principle of parsimony, he cannot therefore evade the question: with what end in view does a particular hypothesis *minimalise human effort*? The question has no singular answer in the domain of astronomy and no singular answer in the domain of psychology. If the record of astronomical discovery has any lesson for the humanistic disciplines, it should surely remind us that many millenia of patient and unpretentious observations on the night sky and the sun's shadow antedated the useful enlistment of higher mathematics in the study of the motion of the celestial bodies. A technique which automatically accelerates the tempo of relevant fact-finding in the formative phase of a science would doubtless be a godsend to the psychologist; but the present impasse does not encourage us to be confident that factor analysis fulfils this role.

The following would thus seem to be a just assessment of the

claims of factor analysis as a so-called tool of research at the present time :

(i) Because its raw data emerge from the domain of *concomitant* variation, factor analysis, like multivariate analysis, is legitimately at best descriptive. Of itself, it can at best lead to a more satisfactory taxonomy, but then only if we are clear about what and why we want to classify. It cannot disclose a causal nexus; and we must judge its usefulness as a means for unmasking unsuspected regularities of nature, of personality or of society by its fruits alone.

(ii) So far the fruitage has been disappointing in the domain of its widest application, and the assertion of its claims in other fields of enquiry may well leave us with the suspicion that its advocates are less reluctant to exploit new markets than to guarantee the intrinsic value of the export.

(iii) At the most elementary level of theory, the invocation of a stochastic model to endorse a recognisable factor-pattern would be unexceptionable if undertaken on the understanding that experiment alone can validify the interpretation suggested by the model; but it is difficult to cite examples of interpretations both consonant with the theory of a suitable stochastic model and also amenable to the verdict of another court of justice.

(iv) Once we abandon the hope of disclosing a unique factor pattern, we can hope to interpret a system of inter-correlations only if content to fall back on an ambiguous principle of parsimony. That any such principle is indeed ambiguous is the sufficient explanation of the existence of different recipes advanced by different schools with no prospect of endorsing the same balance sheet.

(v) Whatsoever be the preferred principle of parsimony, the incorporation of the classical theory of probability in the initial postulates is arbitrary in more ways than one, and involves assumptions which can never be amenable to direct proof. Like Hagen's model of the so-called law of error, any classical model we press into the service of the theory may explain how certain numerical regularities *might* arise but cannot suffice to prove that in fact they *do* so.



PART III

---

*The Calculus of Aggregates*





## MAXWELL AND THE URN OF NATURE

THE ORIGINAL POPULATION of the, now for us notorious, urn of Laplace was a population of *billets*. At what stage the inhabitants of *l'urne que nous interrogeons* became balls the writer has been unable to trace in the literature. Maybe the change occurred *pari passu* with the intrusion of statistical theory into the particulate domain of nineteenth-century experimental science; or possibly it preceded and prepared the way for it. In either event, it signalises the definition of a model peculiarly appropriate to the uses of research when the current conception of atoms and molecules identified them with elastic spheres subject to the Newtonian laws of impact. With this identification before us, we stand on the threshold of what is indisputably to date the most powerful use of the theory of probability; and we shall do well to recall previous remarks on the diversity of its actual or suppositious uses.

In Chapter One we have seen that the word statistics has many uses and that its use in the context of statistical theory alone subsumes four themes which are different at least in the sense that exponents of one or the other respectively operate in watertight compartments of exposition. Thus it is worthy of comment that Kendall's treatise in two volumes on the *Advanced Theory of Statistics* (1943-1946) gives no consideration to the Kinetic Theory of Gases, to Quantum Mechanics or to the Genetical Theory of Populations. It devotes doubtless adequate, perhaps more than adequate, space to Gram-Charlier Series, Tetrachoric functions, Factorial Moments, Tchebychev-Hermite polynomials and Sampling Cumulants of *k*-statistics; but it is silent with respect to Maxwell-Boltzmann, Einstein-Bose and Fermi-Dirac statistics. Indeed, the index cites neither the name of Maxwell nor that of Mendel.

To say this, explicitly carries with it no oblique criticism of Kendall's extremely useful work. I mention these omissions merely to emphasise the fact that a considerable body of specialists in the theory of statistics refrain from disclosing the

relevance of their work to the class of problems here referred to as the *Calculus of Aggregates*. None the less, expositors of the contemporary reorientation of physical concepts unquestionably refer to the construction of hypotheses *en rapport* with recipes which date from Maxwell and Mendel, when they assure us that the statistical is now the canonical formulation of a scientific law. Presumably also theoretical statisticians themselves refer more especially to the domain of such hypotheses when they assure us, if legitimately, that experiment vindicates the success of statistical methods.

Before we can accept the appeal to experience with an easy mind on these terms, we shall need to be clear about the common content of the epithet *statistical* in current usage; and this is now possible only if our reading roams over a wide territory. It is scarcely an exaggeration to say that a course on theoretical statistics delivered to physicists in the department of applied mathematics of any contemporary university would scarcely traverse a single theme dealt with in a course offered to research workers in agricultural science, to students of production engineering or to sociologists. To be sure, the normal distribution would have its niche in both; but the introduction of simple harmonic motion into the treatment of elementary optics, acoustics, the theory of the alternating current and the dynamics of the pendulum does not entitle us to dismiss the obligation to discuss on their own merits problems peculiar to phenomena so diverse. Nor does the use of Fourier series in the treatment of crystal structure, of the conduction of heat and of the diffusion of a solute seduce any reasonable student into believing that the three domains of enquiry have in common anything other than the convenience of exploiting in different situations, and for different reasons, the same algebraic techniques.

If there is indeed any common ground for equating the term statistics as the physicist uses the term when he speaks of Maxwell's Kinetic Theory of Gases as a statistical hypothesis with statistics as R. A. Fisher uses the term in *Statistical Methods for Research Workers*, the student anxious to locate it will get little guidance from any textbooks now in circulation. It is worthy of note that the book last mentioned, though planned

primarily for the use of biologists, does not deal with the genetical theory of populations, an outstanding achievement of the Calculus of Aggregates, the only example of its kind in the terrain of biological research, and a topic to which Fisher himself has contributed. Nor does the same author's *Design of Experiments* set forth the Theory of Error as it emerges in the practice of the observatory and of the physical laboratory.

What I have previously said with reference to Kendall's treatise applies equally to the foregoing remarks. The intention is not to criticise the books here mentioned, because they contain no reference to topics the authors prefer to exclude. I cite such omissions as case material. The naturalist who wishes to understand the contemporary content of the word statistics will wish to know what problems occupy the attention of professional statisticians. If one consults a library with this end in view, one soon becomes aware of an iron curtain between: (a) the Theory of Regression as set forth for the benefit of sociologists, biometricians and educational psychologists; (b) the traditional domain of the *Calculus of Error*, as set forth for astronomers and surveyors. Still more is it true that there is an iron curtain between what I have elsewhere called the *Calculus of Judgments* and the construction of hypotheses which fall within the scope of the *Calculus of Aggregates*.

To define the scope of the latter and the implications of the epithet *statistical* in that context is the theme of this chapter and the next one. It will not be possible to do justice to the task unless we now retrace our steps to the milieu of the Pascal-Fermat correspondence. We must then adjust ourselves to the intellectual climate of a century in which the recognition of the gaseous condition as the third estate of matter was an adventure in wholly uncharted territory. The speculations of Leucippus and Democritus, immortalised in the poem of Lucretius, derided by Aristotle for frivolously fallacious reasons and banned by the Mediaeval Schools on that account, became again widely known in the seventeenth century through the publication of Gassendi's commentaries on Epicurus. The *Commentaries* profoundly influenced the thought of the Newtonian age. The particulate view then re-emerges with the final vin-

dication of the gaseous state of matter as such in the same setting as the exposure of Aristotle's misconceptions about gravity and buoyancy against the background of the way in which the common pump works. Though seventeenth-century thought failed to incorporate a particulate view of matter in a system of quantitative generalisations concerning combination of gases by weight and volume, Hooke himself invoked nitro-aerial particles to offer an explanation of oxidation entirely consistent with the one which Dalton elaborated at the beginning of the nineteenth century.

Gassendi's atoms move in all directions freely. Such motion accounts for the free diffusion of the gaseous state in apparent violation of gravitation. The particles have mass, and particles of like matter are of equal mass. So the density of a gas at a particular pressure depends only on the mean number present in unit space. Being free to move in all directions, they will collide with any partition which obstructs their egress from a vessel. If we move the partition in the direction which lessens the space available to their movements, the number of such collisions will increase. Additional force will thus be necessary to oppose their impact. As its author himself seems to have divined, this brings them into the picture disclosed by Hooke's (so-called Boyle's) law connecting pressure and volume. Daniel Bernoulli, himself a pioneer of stochastic theory in the classical setting, explicitly incorporates the conception in his *Hydrodynamica* (1738).

Before the century ends, we see the stage set for a wider recognition of its usefulness when Charles announced the law connecting the temperature of a gas with its pressure and volume. The work of Black and of James Watt is leading to the recognition of heat as a form of energy. So we can now envisage the possibility of interpreting heat in terms of motion. In everyday life, collision signifies friction, and friction signifies heat. By easy stages we therefore reach the interpretation of temperature in the idiom of *averages*, i.e. in terms of the *mean* incidence of collisions per unit space in unit time, and of conduction as a process of attaining a higher *mean* speed by collision of invading and more swiftly moving particles in a space containing particles in motion at a lower tempo.

If we say that *atomism* is the keynote of naturalistic thought in the nineteenth century, we do not therefore signify that the particulate interpretation of matter was still novel. All we can truly mean is that the particulate view of matter now becomes a background for the construction of hypotheses leading to unique quantitative conclusions susceptible to verification. In the speculations of Dalton, Avogadro, Williamson and Mendelejev the discussion proceeds on the assumption that the gross structure of matter is referable to the *additive* effects of classes of particles *individually alike and endowed with the tangible properties of matter in bulk*. In the intellectual climate at the turn of the half-century, a different picture of the molecule takes shape. We approach its behaviour in the setting of the Gaussian theory of error and its overflow into the domain of vital statistics through the influence of Quetelet's evangel. In short, we identify atoms and molecules with the infinitude of black and white balls in the urn of Nature. The time is ripe for the reception of two generalisations which signalise a new use of the particulate concept and an entirely novel recipe for the prescription of quantitative verifiable hypotheses.

In the background of one we discern a phenomenon which endows the new orientation with a compelling plausibility. Though the germ of the notion is in Gassendi's teaching, the study of Brownian movement in the context of greatly improved microscopic technique during the first half of the nineteenth century made it more easy to visualise the molecules of Avogadro as entities in a state of constant *haphazard* motion. To say as much in the same setting provokes us to seek some connexion between their capricious movements and the concept of randomness implicit in the classical theory of risks in games of chance. The exploitation of this notion explicitly by Clausius (1857) and by Maxwell (1859) necessarily invokes a postulate which is foreign to the Dalton-Mendelejev tradition; and the implications of its invocation still trouble those who demand from science a monolithic statue of nature.

The *individual* particle—atom or molecule—of classical chemistry has such properties of matter in bulk as are relevant to the terms of reference of the theory, and such properties alone. The particles of Clausius and Maxwell are in constant

collision and are therefore losing momentum or gaining it by impact. Their speeds vary *inter se* and that of an individual particle itself varies in less than a twinkling of an eye. If we seek to interpret the gas laws of temperature, pressure and volume in terms of their behaviour, what we can thus postulate about their properties as most relevant to properties of matter in bulk is reducible to averages. Nothing we can know about the *individual* particle is relevant to the behaviour of the aggregate. Though in one sense Maxwell's model of the structure of matter is *en rapport* with a long tradition of thought and with the English tradition of map-making initiated by Gilbert, the relation between the model and the experimental situation has on this account an unfamiliar aspect. If we speak of its essentially novel content as a statistical concept, we can find an intelligible meaning for the otherwise tiresome caption that statistics is the science of averages.

This is not the place in which to do justice to an issue so controversial as the status of the *Principle of Uncertainty* or of *Indeterminacy* in the world-outlook of twentieth-century science, but it may be forgivable if we here pause to remark that certain limitations in what we can say with propriety about individual particles within the framework of a hypothesis stated in such terms are doubtless inherent in our initial assumptions about the model. It is therefore permissible to express a lingering doubt concerning the legitimacy of deriving what spiritual consolations so many contemporary expositors of science do indeed derive from Heisenberg's principle at a macroscopic level before we have conclusively rejected reasonable grounds for supposing that the paradox arises at the level of admissible recipes for constructing hypotheses acceptable or otherwise in virtue of their adequacy in the domain of practice.

Indeed, Jeans (*Dynamical Theory of Gases*) anticipates the paradox as a corollary of the restrictions inherent in the initial assumptions, when he declares:

In the gas of the Kinetic Theory we do not know anything as to the co-ordinates of the individual molecules . . . the problem we have to attack is virtually that of finding as much as we can about the behaviour of a dynamical system without knowing in which of the paths in our generalised space its representative point is moving.

It is the writer's view that the professional logician has the last word in the following passage from John Laird's *Recent Philosophy* (pp. 163-5):

The claim that Heisenberg's "uncertainty principle" knocks the bottom out of determinism seems to be a simple-minded mistake. Everything in nature is what it is, that is, cannot be vague. If precision in the measurement of position is unfriendly towards precision in the measurement of momentum, the trouble lies in the measurement, and is in itself a proof that accuracy of measurement is not the same thing as natural reality unless, indeed, particles do *not* have position, and do *not* have momentum.

In any case, it is an elementary confusion to confound this alleged indefiniteness of nature with "free will," that is, with "indeterminism." The indeterminist holds, say, that he moves his arm freely, but never dreams of denying that his free movements are perfectly definite. What he does deny is that they were inevitably determined by antecedent causes.

Accordingly, if there really is sub-atomic "freewill" quite different arguments must be adduced; and it is plausible to argue, as many modern physicists do, that the macroscopic determinism that astronomers and others assume regarding eclipses and the like does not necessarily imply microscopic determinism, even granting that the macroscopic is composed of the microscopic. For if the macroscopic is a statistical aggregate, it is illogical to apply aggregate-principles forthwith to the components of the aggregate. In life-insurance the death-rate for large numbers is the important matter, and such statistical aggregates do not yield direct information about the chances of survival of some particular insured person.

On the other hand, the difference between aggregates and their components does not even make it plausible to suggest that the former are wholly determined by causes and the latter not at all. There are causes for the death of insured persons (as detectives know) whether or not actuaries concern themselves with any of these particular causes. Again, if the components are determined it is not unreasonable to assume that statistical regularities will continue if no new causes enter, and that they will change if new causes do enter (as the death-rate changes when there is a war). If, however, the components acted quite capriciously, why should there be aggregate constancy?

If and so far as our measurements yield statistical aggregates only we cannot argue that because we know the (macroscopic) past and cannot infer the (microscopic) future, therefore we should

abandon determinism. For, by hypothesis, we do not know the *microscopic* past. Moreover "randomness" is irrelevant. It could be induced in a pack of cards by a shuffling-machine without the faintest denial of determinism. Again, if "randomness" be the opposite of organisation, the human will, being highly organised, ought to be *less* free than most other natural entities.

Improvement of the microscope responsible for the discovery (1827) of Brownian movement also stimulated a new interest in the cellular structure of tissues, the recognition of the role of gametes in sexual generation\* and the study of micro-organisms as agents of fermentation or disease during the last three decades of the first half of the nineteenth century. The discovery of numerical constancy of the chromosomes followed soon after. At more than one level the particulate concept was thus invading biological thought. We encounter it in Darwin's *pangens* incorporated in Galton's erroneous speculations on natural inheritance. What is more important is its role in a now familiar publication which appeared (1866) within seven years of Maxwell's first contribution to the Kinetic Theory of Gases. In this memoir Mendel records no experimental results which signalise an addition to factual knowledge of inheritance based on methods essentially different from those already employed by Thomas Knight, by Gärtner, by Naudin and by his contemporary Laxton during the course of nearly a century of experimental hybridisation. What is noteworthy is a wholly novel interpretation of results substantially identical with those of his predecessors and contemporaries.

The novelty of Mendel's contribution resides partly in the introduction of a particulate concept (designated a *factor* in the days before *Drosophila*, but now the *gene*), and partly in the postulate that fertilisation is comparable to randomwise choice of pairs of balls each member of a pair from one of two urns, though the author does not explicitly exploit the metaphor. At the most elementary level—unit character differences in the idiom of the childhood of modern genetics—we may conceive the balls in the urn to be of two sorts: black and white. We identify those of one urn as male gametes, those

\* Brown was a botanist. Amici (1821) who first recorded the fertilisation of the ovule by a single pollen grain was a physicist.



of the other as female. We score the pair as a success if black-black or black-white and as a failure if white-white is the zygote. Mendel carried this conception further to interpret with equal acceptability the outcome of his own crosses involving two so-called unit character differences. We then conceive that each urn contains balls of four sorts and adapt our scoring system accordingly.

Unlike that of Maxwell, Mendel's theory failed to attract recognition from his contemporaries. When it engaged the attention of a later generation several circumstances were propitious to its reception. One of these recalls the part played by the study of Brownian movement in the background of Maxwell's own innovation. Mendel knew that only one pollen grain unites with the ovum of the seed plant; but the essentials of animal breeding were still obscure and controversial in his own time. In 1875 Hertwig and Fol first observed the fertilisation of the egg of an Echinoderm. This observation brought the study of animal inheritance within the terms of reference of Mendel's hypothesis, but it is here of special interest for another reason. That only one sperm penetrates the egg is not the only noteworthy feature of a process we now regard as the keystone of bisexual inheritance. What is also remarkable is the fact that spermatozoa execute movements hither and thither in all directions like the movements of Brownian particles. Seemingly reasonable betting odds on the successful attainment of its goal are the same for any one sperm as for any other. All that we can see through the microscope is thus consistent with the conclusion that a drop of seminal fluid is one of nature's urns.

The reader should not conclude that such new knowledge of the material basis of inheritance led to any immediate and explicit formulation of hereditary transmission in terms of a stochastic model. What makes the Mendelian hypothesis of special relevance to our attempt to clarify the diverse current meanings of such words as probability and statistics is that Mendel himself never explicitly invoked the classical doctrine, and his foremost expositors developed the theory for more than a decade in opposition to Pearson's assertion of the claims of statistical methods without realising that the theory of the gene

is a statistical hypothesis in the same sense that the Kinetic Theory of Gases is also a statistical hypothesis. That the theory of the gene did so develop without explicit recognition of the formal identity of the operations invoked and the basic theorems of the classical theory is instructive for reasons we shall examine in the next chapter. Here it suffices to indicate one relevant difference between the problems which Maxwell and Mendel respectively assailed.

In the initial stages of the growth of what we now call the theory of the gene, the main preoccupation of the investigator in the field was with clear-cut qualitative differences between varieties with a view to testing the relevance of the principle of segregation for a wide range of structural features definable in such terms and for a wide range of species, plant or animal. In such enquiries, the method of scoring is that of simple enumeration, and it is possible to exploit the initial assumptions by recourse to simple algorithms without importing a tailor-made algebra to accomplish the end in view. From one point of view, this simplicity of the formal apparatus gives the theory a unique interest in the context of our theme. For we can explore the implications of the use of the word *statistical* in a calculus of aggregates without the additional distraction of traversing a formidable terrain of mathematical operations. On that account I shall devote the next chapter to the theory of the gene as case material.

From another viewpoint, the compelling simplicity of the initial assumptions we make in the theory of the gene has a drawback. When the scope of the theory enlarges, and we then feel the need to interpret our algorithms in the idiom of the classical theory of probability, we do so with no disposition to examine our preference for the stochastic model we enlist. Nowadays, every biological student who attends a vacation course at Woods Hole, at Plymouth or at Naples, sees with his or her own eyes what Hertwig and Fol saw, as indeed did all the author's first-year students in the University of Cape Town. When we translate Mendel's interpretation into the idiom of the urn model, we do so effortlessly with little disposition to ask why we have chosen a model with the properties peculiar to it.

Our approach to the sister theory is necessarily different. Maxwell's concern was to explore the implications of the movements of particles *vis-à-vis* the relation of the gas laws to the physical phenomena contingent on chemical reaction involving molecular recombinations. Initially, then, we are dealing with a system of scores referable to speeds. We are thus in the domain of representative scoring, and indeed of distributions which admit of an infinitude of score values. All our particles—in a homogeneous gas—have the same mass, but their speeds are variable in virtue of the collisions. The initial problem is to define what we can say about the distribution of particle speed referable to a particular gas at a particular temperature and pressure within the framework of the assumption that the particles behave like elastic spheres subject to the Newtonian laws of impact.

In his memorable paper delivered to the British Association in Aberdeen, Maxwell (1857) attacked this problem by assuming that the molecules of a gas at a given instant may be anywhere. He endows the velocity of each particle with three Cartesian co-ordinates ( $u, v, w$ ) to each of which he assigns a function which specifies a frequency definitive of its particular value referable to an arbitrary origin. He then cites the following conclusion. If  $f(u)$  is the relative frequency of particles with speed  $u$  after unit interval of time in the appropriate axis of the Cartesian space:

$$f(u) = Ce^{-Ku^2} \quad \text{and} \quad f(u,v,w) = C^3e^{-K(u^2+v^2+w^2)}$$

This relation is the foundation-stone of the theory associated especially with the names of Maxwell and Boltzmann; but the original publication does not make explicit all the steps in the reasoning which led Maxwell to advance it. Its importance resides in the fact that it embodies a *principle of equilibrium*, i.e. it asserts, without invoking any restriction on the initial state, what proportion of particles are moving at the end of a unit interval of time with an assigned velocity in a particular direction specified by the elementary rule for composition of velocities expressed in terms of Cartesian components. It is scarcely necessary to point out that this principle has a familiar aspect. The algebraic function which it embodies is one which

had emerged from the soil of the classical theory of probability in the writings of Laplace. Hagen had dramatised its convenience in a domain of models conceived in terms strictly consistent with the classical theory.

Thus Maxwell had to hand a system of formal algebra uniquely associated with the term probability as used by the physicists of his time; whence his much-quoted aphorism "the true logic for this world is the calculus of probability, the only mathematics for practical men." A saying so strongly entrenched by oft-repetition would be less exceptionable if we could justifiably equate the practical man with the physicist of Maxwell's time. Even so, the form of words is mischievous, because most practical men, including many physicists, use the word probability, if at all, to describe a personal sentiment to which considerations suggesting the construction of models in the Hagen tradition have no compelling relevance.

We have had occasion to note that Hagen's choice of a classical model was quite arbitrary. It was not obvious in Maxwell's time that an explicit formulation of a model which incorporates all the assumptions appropriate to the dynamics of a system of particles is equally arbitrary, though the outcome has more comprehensive claims to usefulness. Few, if any, would now assert that Maxwell's formulation of the distribution law referred to above is a satisfactory proof in this sense. Indeed, a single citation from Jeans (*Dynamical Theory of Gases*, pp. 56-7) will suffice to disclose a lack of unanimity about the validity of the postulates he and others have invoked. Referring to objections w.r.t. his original deductive statement of the theory, Jeans remarks:

This proof must be admitted to be unsatisfactory, because it *assumes* the three velocity components to be independent. The velocities do not, however, enter independently into the dynamical equations of collisions between molecules, so that until the contrary has been proved, we should expect to find correlation between these velocities.

On account of this defect, Maxwell attempted a second proof, which after emendations by Boltzmann and Lorentz assumes the form given in Chapter II. It is, however, very doubtful whether this proof can claim any superiority on grounds of logical consistency

or completeness over Maxwell's original proof. The later proof finds it necessary to assume that there is no correlation between the velocity and space co-ordinates, while the earlier proof merely assumed that there was no correlation between the separate velocity components *inter se*. In each case the dynamical conditions equally suggest correlation until the contrary has been proved, and it would be hard to give reasons why one assumption of no correlation is more justifiable than the other. It should be mentioned that Burbury was always of opinion that the later proof of Maxwell was not only logically unsound, but led to an inaccurate result. . . .

A second class of proof of the law is represented by the proof which has been given in this chapter. . . . As important examples of this class of proof may be mentioned a proof due to Kirchhoff, given in his lectures, and one due to Meyer and Pirogoff, given in Meyer's *Kinetic Theory of Gases*. Both these proofs are found on analysis to depend upon a use of the calculus of probabilities which cannot be justified. The proof given in this chapter is my own: it also has been criticised by Burbury.

It is not here necessary to examine such subsequent attempts to provide a better justification of the normal law of molecular speed components than that of Maxwell himself. By enlisting the same type of reasoning to describe the behaviour of a new battery of particles, more recent developments of physical theory have brought into focus more clearly than hitherto what assumptions we then have to make. The dilemma arising from one such set of assumptions is easy to grasp without recourse to dynamical concepts. If we speak of a randomwise distribution of particles in space, we imply that we assign equal probability to each of all possible configurations appropriately specified. This leaves open the question: how shall we specify such configurations? If we then conceive our space as a grid of cells to each of which we can allocate one or more balls, we may seek in more than one way an answer to the question at the level of stochastic theory conceived in terms of proportionate choice.

Given a grid of  $n$  cells, we may specify the number of ways of allocating  $r$  balls to the cells in accordance with different suppositions of which only 3 need concern us:

- (a) the balls are distinguishable, and there is no restriction on the number 0, 1 . . .  $r$  allocated to one and the same cell;

(b) the balls are indistinguishable, and there is no restriction on the number per cell as before;

(c) the balls are indistinguishable, and one cell can accommodate only one ball.

If our problem is (a), the number of different allocations is  $n^r$  since we can put the first ball in any one of  $n$  cells and so forth. In the formal definition (p. 41) of algebraic probability in terms of proportionate choice, we shall say that any one of the  $n^r$  different patterns of allocation has therefore a probability  $n^{-r}$ . Now each of these  $n^r$  different configurations is one of a class of different linear arrangements of balls in the individual cells of one and the same cell-set in the same way. All members of such a class would be indistinguishable, if the balls were alike as in (b); but we cannot assign to every such distinguishable group an equal probability, without abandoning the foregoing specification of the probability assigned to each member ( $n^{-r}$ ), since different groups will contain a different number of members. Thus the class of allocations specified by the fact that the  $h$ th cell contains  $r$  balls consists of one member. The class specified by the allocation of 1 ball each to the first  $r$  cells consists of  $r!$  members. For the 7-cell and 6-ball set-up specified by the class represented in terms of balls (B) per cell and empty cells (O) as BOBBBOOBBO, the number of different members is  $6! \div (1!3!2!) = 60$ , if each ball is distinguishable and the probability of getting such a class specification is therefore  $60 \cdot 7^{-6} \simeq \frac{1}{2000}$ .

Specification of all possible ways of making the second type of allocation, (b) above, is not so obvious, unless we think of the  $n$ -fold grid laid out in a row, with one or more balls (B) in a cell separated from its neighbour by a partition (P), some cells being empty (O). For a grid of 8 cells to which we allocate balls, our pattern of one such allocation will then be:

P O P B B P O P B B B P O P O P B P O P

Only  $7 = (n - 1)$  partitions separate adjacent cells, and our problem is thus to state how many linear arrangements of  $(n - 1)$  partitions and  $r$  balls are possible. This is the familiar problem of the number of permutations of  $m = (n + r - 1)$

things all taken,  $r$  being alike of one kind and  $(n - 1)$  alike of another, i.e.  $(n + r - 1)! \div (n - 1)! \cdot r!$ . In the sense defined above, we may then assign to any of the allocations specified by (b) the probability:

$$P_b = \frac{(n - 1)! r!}{(n + r - 1)!} \quad . \quad . \quad . \quad . \quad (i)$$

Our third type of allocation offers no difficulty, if we first assume that the balls are distinguishable. Then we can allocate one of the  $r$  to any of  $n$  cells, one of the remaining  $(r - 1)$  to any one of  $(n - 1)$  cells and so on. There will thus be  $n^{(r)}$  different ways of making the allocation, but each such allocation involves  $r!$  different linear arrangements of  $r$  objects taken all at a time. If the balls are indistinguishable, the number of distinguishable allocations will therefore be  $n^{(r)} \div r!$ . We may accordingly assign to any one of the allocations defined by (c) above a probability:

$$P_c = \frac{r!(n - r)!}{n!} \quad . \quad . \quad . \quad . \quad (ii)$$

Whereas we have seen that the specification of the probability of the particular allocation for a set-up of 7 cells and 6 balls as BOBBBOOBBO in accordance with (a) is approximately  $\frac{1}{2000}$ , we obtain from the definition of the event in accordance with (b) from (i) and in accordance with (c) from (ii)

$$(b) \frac{6! \cdot 6!}{12!} = \frac{1}{924} \quad ; \quad (c) \frac{6! \cdot 1!}{7!} = \frac{1}{7}$$

The three foregoing methods of scoring the event of allocating  $r$  balls to  $n$  cells of a grid as a prerequisite to definition of the probability of a specified grid distribution of the balls correspond to three different systems of scoring the position of a system of particles in space; and the initial choice of the scoring system leads to different formulations referred to respectively as: (a) Maxwell-Boltzmann statistics; (b) Einstein-Bose statistics; (c) Fermi-Dirac statistics mentioned on page 279. The formulations lead to different results. Einstein-Bose statistics give a satisfactory account of photons, nuclei and atoms with an even number of elementary particles. Fermi-

Dirac statistics lead to a satisfactory description of phenomena involving electrons, protons and neutrons. Maxwell-Boltzmann statistics are not wholly satisfactory for particles of any sort. "We have here," says Feller rightly, "an instructive example of the impossibility of selecting or justifying probability models by *a priori* arguments. In fact, *no pure reasoning* could tell us that photons and protons could not obey the same probability laws."

In the next chapter we shall examine the Mendelian theory of populations more fully to clarify the terms of reference of a statistical hypothesis as we use the term *statistical* in the experimental sciences without blurring semantic issues by introducing kinematical algebra irrelevant to the stochastic assumptions. At this stage it is already permissible to anticipate the main conclusions which emerge against the background of the foregoing historical narration. Accordingly, we shall define a statistical hypothesis in the framework of the Calculus of Aggregates as a hypothesis which explicitly invokes, or is interpretable in terms of a model of the die, urn, card pack or lottery wheel type in the sense that its properties are not specifiable in terms of the unique properties of a particular face, ball, card or sector. Though we may make certain initial assertions, e.g. about the shape or mass of all our particles, as we may make certain assertions about the colour or shape of balls in an urn, we shall otherwise confine ourselves to statements about the distribution of the scores assigned to the components on the assumption that such statements hold good, if, and only if, the sample is very large. When we interpret the behaviour of matter in the bulk in terms of particles outside the range of immediate inspection, the assumption last stated is trivial. For we are clear enough about the end in view. On this understanding, we may make the following assertions:

- (i) the choice of a stochastic model endorses reliance on the algebraic theory of probability only in so far as we deem the algebraic theory of probability to describe its relevant properties in terms of observable external events;
- (ii) the choice of any such model is arbitrary, as is the choice of any non-stochastic model which physicists of the



seventeenth and eighteenth century invoked, e.g. waves on the surface of a pond or elastic springs in a viscous jelly;

(iii) what justification we deem to be adequate for such a choice is the same as adequate justification for preferring any one non-stochastic model to any other, i.e. the outcome is *both* capable of accommodating what we know about a certain range of phenomena *and* of pointing to new hitherto unsuspected regularities of nature.

From the foregoing summary two other conclusions are inescapable:

(a) what we do when we invoke the formal algebra of the classical theory has at no stage from the initial step in the construction of a statistical hypothesis to its final confirmation any necessary relevance to a concept of probability defined as the measure of our legitimate conviction concerning the truth of a proposition;

(b) what we mean by justifying the choice of an appropriate model is neither more nor less than what we mean by confirming the hypothesis itself.

The first of the two conclusions last stated signifies that we cannot legitimately invoke the success of the foregoing recipe for law-making as a justification for faith in a calculus of judgments conceived in terms of a definition of probability referable to legitimate intensity of conviction, unless we also define the word legitimate in a particular way. We must then define it in the strictly behaviourist framework of verbo-factual parallelism. Our definition of probability as a measure of our legitimate conviction that a statement is correct will then convey no more than the assertion that: (i) we make a statement in conformity with a prescribed rule; (ii) an assignable proportion of our statements will be correct in the long run if we adhere to the rule consistently.

If we restrict the proper domain of the calculus of probability to a particular class of external events, i.e. long-run score frequencies definitive of the behaviour of particular model situations, we are using the same language in the domains I have distinguished at the outset as a *Calculus of Judgments* and as a *Calculus of Aggregates*; but the terms on which we invoke the

formal algebra of the classical theory or the model to which we deem it to be appropriate will still be different. In the domain of the latter we justify an arbitrary prior choice of the model by whether it works; and a single *experimentum crucis* suffices to relegate it to the limbo of profitless speculation, if it does not work. In the domain of the former we can endorse the rule we operate only by correct prior choice of the model. We can entertain no hope of justification by works, because the rule permits us to make a mistake on any single occasion, and the criterion of its relevance must be the outcome of an *endless* sequence of experiments. We must then concede to Bacon that radical errors in the first concoction cannot be mended by subsequent remedies, however excellent.

## CHAPTER THIRTEEN

# MENDELISM AND THE TWO MEANINGS OF PROBABILITY

THOUGH WE NOW recognise Mendel's theory and the superstructure raised on its foundations as an outstanding example of a statistical hypothesis in the same sense that Maxwell's Kinetic Theory of Gases is a statistical hypothesis, no one would have suggested the inclusion of a course on the theory of probability in a curriculum of biological instruction prerequisite to the study of experimental genetics in 1913, when the writer attended Punnet's course at Cambridge—in that year the only one of its kind in Britain. The early stages of testing and extending Mendel's hypothesis after the independent rediscovery of the principle by Tschermak, by Correns and by de Vries (1900), followed shortly by the work of Bateson and of Cuenot (1902) who independently demonstrated the applicability of the principle of segregation to bisexual inheritance of animals, recall a story in the memoirs of Sherlock Holmes about the dog that barked in the night. The clue is (my dear Watson) that the dog did not bark.

As I have elsewhere said, our more conciliatory instructors in such days conceded the existence of two sorts of inheritance—statistical or *regression* and experimental or *alternating*. Otherwise, they advised us, and with good reason, that statistics had done a lot of harm. Should any reader of this book regard my *commentar* in Chapters Seven to Nine as irresponsibly facetious or as unjust to the memory of the redoubtable founders of the Eugenic cult, a citation from Bateson's book *Mendel's Principles of Heredity* (1913) will dispel the illusion. In the passage which follows we see the Galton-Pearson episode from the viewpoint of the outstanding pioneer of genetics as an experimental science in Britain:

Of the so-called investigations of heredity pursued by extensions of Galton's non-analytical method and promoted by Professor Pearson and the English Biometrical school it is now scarcely

necessary to speak. That such work may ultimately contribute to the development of statistical theory cannot be denied, but as applied to the problems of heredity the effort has resulted only in the concealment of that order which it was ostensibly undertaken to reveal. A preliminary acquaintance with the natural history of heredity and variation was sufficient to throw doubt on the foundations of these elaborate researches. *To those who hereafter may study this episode in the history of biological science it will appear inexplicable that work so unsound in construction should have been respectfully received by the scientific world.* (Italics inserted.)

The index of Bateson's noteworthy exposition in which this passage occurs does not list the word probability or the word chance; and I have not been able to find either in the text. The following citation in which the idiom of *averages* is consistent with that of the author's argument throughout does indeed contain an explicit reference to *chance* in the English translation of Mendel's own original memoir; but it scarcely justifies the assumption that he used it with any awareness of its meaning in the context of the classical calculus:

In individual flowers and in individual plants, however, the ratios in which the forms of the series are produced may suffer not inconsiderable fluctuations. Apart from the fact that the numbers in which both sorts of egg cells occur in the seed vessels can only be regarded as equal on the average, it remains purely a *matter of chance* which of the two sorts of pollen may fertilise each separate egg cell. For this reason the separate values must necessarily be subject to fluctuations, and there are even extreme cases possible, as were described earlier in connection with the experiments on the form of the seed and the colour of the albumen. The true ratios of the numbers can only be ascertained by an average reduced from the sum of as many single values as possible; the greater the number the more are merely *chance* effects eliminated. (Italics inserted.)

We may seek an explanation of this reticence at more than one level. One, which has been the subject of comment *en passant* in Chapter Ten, is that the convenience of an algebraic formulation of the theory of the relevant models does not become imperative till: (a) the theory of hybridisation takes within its scope quantitative differences referable to many gene substitutions: (b) the genetical theory of populations advances

beyond the level at which Mendel left it. Another relevant circumstance emerges from one of the foregoing citations. During the first two decades of the Mendelian renaissance, biologists in the following of Bateson and Morgan were mostly familiar with the calculus of probability, if at all, through the writings of Pearson, himself a staunch advocate of the backward look and of the identification of the concept of probability with a state of mind. For one reason or another, the pioneers of the Mendelian renaissance were one and all content, and well content, to interpret their findings by recourse to simple algorithms to which the Pearson-Laplace doctrine of inverse probability has no relevance whatsoever.

The bearing of these reflections on the present crisis of theoretical statistics will become more apparent, if we now set forth Mendel's hypothesis first in the idiom of the classical theory and then in the idiom of his earliest expositors. With that end in view, we may conveniently distinguish two levels of discussion in Mendel's own memoir. At one, he deals with the interpretation of his own experiments on what Knight had called in the closing years of the eighteenth century the *Law of the Splitting of Hybrids*. At the other, he asks and correctly answers the question: how does it come about that close inbreeding results in producing the pure stocks necessary for the conduct of such experiments? His approach to the first issue suffices to specify a stochastic model for the theory of the gene conceived in structural terms. I shall call this the *First Order Model*. His approach to the second foreshadows models of a different order as a basis for the genetical theory of populations. I shall call these *Models of the Second and of the Third Order*.

*The First Order Model.* Our first-order model will be two types of urn (A and B), the contents of which we identify respectively with the gametes of the father (A) and of the mother (B). Each urn contains an infinitude of balls, and we choose random-wise one ball from one of each type of urn. We identify the 2-fold sample (pair chosen) as a zygote. If we now appropriately prescribe the contents of each urn, we have all the data for specifying the relevant constitution of a zygote in terms consistent with the eighteenth-century theory of probability. That is to say, we can state what proportion of zygotes

will have the specified constitution, if we continue the sampling process long enough, in other words, if we rear a large enough progeny of two parents with the prescribed equipment of genes.

To accommodate all possibilities which arise when our concern is with only *one* gene substitution, we shall admit two ways of specifying the urn contents consistent with normal chromosome behaviour:

(a) each urn may contain white balls only, black balls only or black and white in equal numbers (*autosomal* case);

(b) urn B may contain white balls only, black balls only or white and black in equal numbers, but urn A must contain either white balls and grey balls in equal numbers or black and grey balls in equal numbers (*X-linked* case).

A single example will suffice to illustrate how we accommodate situations involving more than one gene substitution. If we put two autosomal gene substitutions in the picture, we shall have to admit the possibility that both our urns contain balls of 1, 2, 3 or 4 sorts (blue, green, yellow and red). Mendel's own experiments at this level constitute a special case of a more general pattern which will emerge if we now specify the probabilities appropriate to the relevant unit sample distribution of the urn.

If we denote by  $p$  the probability of drawing a white ball and  $q = (1 - p)$  that of drawing a black ball from either urn of (a) above, the admissible values of  $p$  are 0,  $\frac{1}{2}$ , 1 for all possible specifications of either A or B appropriate to the interpretation of Mendel's monohybrid experiments. When we also allow for the possibility (b) of sex-linked inheritance  $p = \frac{1}{2}$  or 0 for urn A and 0,  $\frac{1}{2}$  and 1 for urn B. We have then completely specified the problem for a given pair of parents or for parents of specified genotypes.

When there are two autosomal gene substitutions to consider, we shall have to specify the probabilities of drawing a blue ( $AB$ ), green ( $Ab$ ), yellow ( $aB$ ) or red ( $ab$ ) ball respectively by  $p, q, r$  and  $s = (1 - p - q - r)$  with the restriction that  $(p + q)$  can have values 0,  $\frac{1}{2}$  or 1 only and  $(r + s)$  can have values of 0,  $\frac{1}{2}$  or 1 only as prescribed by the constitution of the urn, i.e. parental genotype. For the most generalised situation disclosed

by the elucidation of linkage either  $(p + s) = c$  or  $(q + r) = c$  in which  $c \leq \frac{1}{2}$  is a constant for the particular pair of gene substitutions. For the particular case  $c = \frac{1}{2}$  we have  $p = q = r = s = \frac{1}{4}$ . This is a formal statement of Mendel's so-called second law.

To proceed, we have merely to agree about the specification of  $p$ ,  $q$ , etc., and how to score the two-fold sample. We may then subsume as follows all the results of Mendel's monohybrid experiment by the following frequency distribution of pairs of white balls (WW), mixed balls (BW) and pairs of black balls (BB), if we distinguish  $p_a$  and  $p_b$  as the proportions of white balls in urns of type A and type B respectively:

WW	BW	BB
$p_a p_b$	$p_a q_b + p_b q_a$	$q_a q_b$

Since this expression is symmetrical with respect to  $p_a = (1 - q_a)$  and  $p_b = (1 - q_b)$ , six situations arise. We may explicitly identify them by classifying our genotypes as R = WW, H = BW and D = BB, the matings then being specified uniquely by the numerical values of  $p_a$  and  $p_b$ :

RR	$p_a = 1 = p_b$
RH or HR	$p_a = 1 ; p_b = \frac{1}{2}$ or $p_a = \frac{1}{2} ; p_b = 1$
RD or DR	$p_a = 1 ; p_b = 0$ or $p_a = 0 ; p_b = 1$
HH	$p_a = \frac{1}{2} = p_b$
HD or DH	$p_a = \frac{1}{2} ; p_b = 0$ or $p_a = 0 ; p_b = \frac{1}{2}$
DD	$p_a = 0 = p_b$

The foregoing define all possible constitutions of the urn model of our mating system, and the general expression cited above then yields the following in agreement with Mendel's results:

	Offspring:	R	H	D
<i>Mating</i>				
RR		1	0	0
RH or HR		$\frac{1}{2}$	$\frac{1}{2}$	0
RD or DR		0	1	0
HH		$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
HD or DH		0	$\frac{1}{2}$	$\frac{1}{2}$
DD		0	0	1

Other possibilities are obtainable in the same way; and the only necessary elaboration of the foregoing model to accommodate the theory of the gene in the absence of mutation arises when there is an abnormal chromosome complex, e.g. a cross between heterozygous eyeless ( $E E e$ ) of triploid stock and the normal heterozygote ( $E e$ ). Here  $p = \frac{1}{3}$  for one urn and  $p = \frac{1}{2}$  for the other. Thus the 2-fold sample frequency of the class eyeless ( $ee$ ) is  $\frac{1}{6}$ , a ratio of 5 dominants to 1 recessive.

*The Model of the Second and Third Order.* In the domain of the model of the first order we ask the following question: if we can specify the genotypes of the parents, what is the frequency distribution of genotypes referable to their offspring? In the domain of a class of models we shall now deal with in less detail we ask: if we can specify the frequency distribution of genotypes in a generation of parents, what is the corresponding distribution referable to their offspring? We can make the last question meaningful only if we can also specify the mating system, i.e. the relevant conditions which prescribe how a male parent of a specified genotype mates with a female parent of a specified genotype. The outcome is the way the genotype structure of a population changes if we impose on it a particular mating system or selection. For any specified genotype, we express this by citing the general  $(n + r)$ th term of a series exhibiting its frequency in the  $(n + r)$ th generation in terms of its initial frequency, that of the  $n$ th. The necessary preliminaries may end, if we can express the genotypic frequencies of one generation uniquely in terms of those of its predecessor, in which case we shall require what we shall define as a model of the *second order*. It may be necessary to specify the grand-parental, as well as the parental, generation in which event we shall encounter *recurrent series*, and our model will be a *third-order model*.

For present purposes it will suffice, if we illustrate the specification of a second order model by reference to the particular situation which arises when our concern is with only one autosomal gene substitution. We shall then postulate:

- (a) a pack containing  $N$  cards of nine types marked as such by the letters A, B, C . . . I, the numbers of cards of the



several types being respectively  $a_0N$ ,  $b_0N$  . . .  $i_0N$ , so that  $a_0$  is the probability of drawing a card of type A,  $b_0$  of drawing a card of type B, and so on, in a single trial;

(b) a series of urns likewise labelled A-I, each constituted as follows:

- A            white balls only;
- B and C    white and grey balls in equal numbers;
- D and E    grey balls only;
- F            white, grey and black balls in the ratio 1 : 2 : 1;
- G and H    grey and black balls in equal numbers;
- I            black balls only;

The rules of the game are as follows:

(a) at each trial, draw one card from the pack, record its type, replace and reshuffle;

(b) draw *one* ball from the urn bearing the label corresponding to the type of card chosen;

(c) record the result as W, G or B in terms of the specification of the ball chosen.

Here the nine card types (A-I) stand for matings: RR, RH, HR, RD, DR, HH, HD, DH and DD respectively. Accordingly, the parental ( $n$ th) generation has the following genotypic frequency distribution:

$$\left. \begin{aligned} R_n &= a_0 + \frac{1}{2}b_0 + \frac{1}{2}c_0 + \frac{1}{2}d_0 + \frac{1}{2}e_0 = u_n \\ H_n &= f_0 + \frac{1}{2}b_0 + \frac{1}{2}c_0 + \frac{1}{2}g_0 + \frac{1}{2}h_0 = 2v_n \\ D_n &= i_0 + \frac{1}{2}d_0 + \frac{1}{2}e_0 + \frac{1}{2}g_0 + \frac{1}{2}h_0 = w_n \end{aligned} \right\} \quad (\text{ii})$$

By definition  $(u_n + 2v_n + w_n) = 1$ , in the foregoing. We interpret our three classes of balls as the three genotypic classes of the offspring of the matings, viz. white as R, grey as H and black as D. We can then specify their frequencies by recourse to the multiplicative rule. Thus the probability that we shall sample in urn F is  $f_0$  and the unconditional probabilities of getting a white, grey or black ball from urn F are  $\frac{1}{4}f_0$ ,  $\frac{1}{2}f_0$  and  $\frac{1}{4}f_0$  respectively. We may then audit the complete balance sheet

of the 2-stage sampling process by recourse to the addition theorem:

Mating	Offspring		
	R	H	D
RR	$a_0$	o	o
RH	$\frac{1}{2}b_0$	$\frac{1}{2}b_0$	o
HR	$\frac{1}{2}c_0$	$\frac{1}{2}c_0$	o
RD	o	$d_0$	o
DR	o	$e_0$	o
HH	$\frac{1}{4}f_0$	$\frac{1}{2}f_0$	$\frac{1}{4}f_0$
HD	o	$\frac{1}{2}g_0$	$\frac{1}{2}g_0$
DH	o	$\frac{1}{2}h_0$	$\frac{1}{2}h_0$
DD	o	o	$i_0$

Thus the genotypic frequency distribution (*g.f.d*) of the generation of offspring is:

$$\left. \begin{aligned} R_{n+1} &= a_0 + \frac{1}{2}b_0 + \frac{1}{2}c_0 + \frac{1}{4}f_0 = u_{n+1} \\ H_{n+1} &= d_0 + e_0 + \frac{1}{2}b_0 + \frac{1}{2}c_0 + \frac{1}{2}f_0 + \frac{1}{2}g_0 + \frac{1}{2}h_0 = 2v_{n+1} \\ D_{n+1} &= i_0 + \frac{1}{2}h_0 + \frac{1}{2}g_0 + \frac{1}{4}f_0 = w_{n+1} \end{aligned} \right\} \text{(iii)}$$

Here by definition  $(u_{n+1} + 2v_{n+1} + w_{n+1}) = 1$ . By giving  $a_0, b_0, c_0$ , etc., appropriate values we can now define the results of several systems of mating. For illustrative purposes, two will suffice.

(i) There is no restriction on the randomwise choice of parents, i.e. mating is *non-assortative*. For a parental distribution of  $R_n, H_n, D_n$ , specified as  $u_n, 2v_n$  and  $w_n$  in (ii) above, the probabilities of mating  $R_nR_n, R_nH_n, H_nR_n$ , etc., are  $u_n^2, 2u_nv_n, 2u_nv_n$ , etc., i.e.:

$$\begin{aligned} a_0 &= u_n^2 ; b_0 = 2u_nv_n ; c_0 = 2u_nv_n \\ d_0 &= u_nw_n ; e_0 = u_nw_n ; f_0 = 4v_n^2 \\ g_0 &= 2v_nw_n ; h_0 = 2v_nw_n ; i_0 = w_n^2 \end{aligned}$$

If we now paint these values into (iii) we get:

$$\left. \begin{aligned} u_{n+1} &= u_n^2 + 2u_nv_n + v_n^2 = (u_n + v_n)^2 \\ 2v_{n+1} &= 2u_nw_n + 2u_nv_n + 2v_nw_n + 2v_n^2 \\ &= 2(u_n + v_n)(v_n + w_n) \\ w_{n+1} &= w_n^2 + 2v_nw_n + v_n^2 = (v_n + w_n)^2 \end{aligned} \right\} \quad (\text{iv})$$

If these offspring mate randomwise without restriction, we start a new cycle and must accordingly redefine the constitution of our new card pack as:

$$\begin{aligned} a_1 &= (u_n + v_n)^4 ; \quad b_1 = 2(u_n + v_n)^3(v_n + w_n) = c_1 \\ d_1 &= (u_n + v_n)^2(v_n + w_n)^2 = e_1 ; \\ f_1 &= 4(u_n + v_n)^2(v_n + w_n)^2 \\ g_1 &= 2(u_n + v_n)(v_n + w_n)^3 ; \quad i_1 = (v_n + w_n)^4 \end{aligned}$$

The definition of the g.f.d. for the offspring will then be in accordance with (iii), e.g.:

$$\begin{aligned} u_{n+2} &= a_1 + \frac{1}{2}b_1 + \frac{1}{2}c_1 + \frac{1}{4}f_1 \\ &= (u_n + v_n)^4 + 2(u_n + v_n)^3(v_n + w_n) \\ &\quad + (u_n + v_n)^2(v_n + w_n)^2 \\ &= (u_n + v_n)^2[(u_n + v_n)^2 \\ &\quad + 2(u_n + v_n)(v_n + w_n) + (v_n + w_n)^2] \\ &= (u_n + v_n)^2(u_n + 2v_n + w_n)^2 \\ &= (u_n + v_n)^2 \end{aligned}$$

Whence from (iv):

$$u_{n+2} = u_{n+1}$$

Similarly  $v_{n+2} = v_{n+1}$  and  $w_{n+2} = w_{n+1}$ . Thus the genetic structure of the population attains a constant state after one generation of mating randomwise without restriction. We express this by saying that non-assortative mating in a population classified w.r.t. a single autosomal gene substitution attains equilibrium in one generation.

(ii) Let us now suppose that only like genotypes can mate *inter se* as is true if there is only *self-fertilisation* or if there is *complete positive assortative mating without dominance*. We then

admit only matings RR, HH and DD so that  $b_o = c_o = d_o = e_o = g_o = h_o = 0$ . Whence we obtain from (ii) and (iii) at one step the genotypic frequency distributions:

	R	H	D	
Parents from (ii)	$a_o$	$f_o$	$i_o$	(v)
Offspring from (iii)	$(a_o + \frac{1}{4}f_o)$	$\frac{1}{2}f_o$	$(i_o + \frac{1}{4}f_o)$	

If we now repeat the process by reconstituting our card pack accordingly as in the previous example, we get:

$$\begin{array}{ccc} R_{n+2} & H_{n+2} & D_{n+2} \\ a_o + \frac{1}{2}f_o & \frac{1}{4}f_o & i_o + \frac{1}{2}f_o \end{array}$$

Thus the proportion of heterozygotes falls off 50 per cent in each generation, a result Mendel himself gives, i.e.:

$$H_{n+1} = \frac{1}{2}H_n ; H_{n+2} = \frac{1}{4}H_n ;$$

$$H_{n+3} = \frac{1}{8}H_n ; \dots H_{n+r} = 2^{-r} \cdot H_n$$

\* \* \* \*

The foregoing examples do not exhaust the uses of the second order model; but it may be necessary to introduce a 3-stage sampling programme, in which event we arrive at recurrent series such as (v) involving two antecedent terms, i.e.:

$$S_{n+2} = a \cdot S_{n+1} + b \cdot S_n$$

A single example will suffice. If we follow a system of inbreeding by brother-sister mating in a population classified w.r.t. a single autosomal gene substitution, all mating sub-populations will be:

- |                          |   |
|--------------------------|---|
| I For parents RR         | RR  |
| II For parents RH or HR  | $\frac{1}{4}RR ; \frac{1}{2}RH \text{ or } HR ; \frac{1}{4}HH$  |
| III For parents RD or DR | HH  |
| IV For parents HH        | $\frac{1}{16}RR ; \frac{1}{4}RH \text{ or } HR ; \frac{1}{8}RD \text{ or } DR ;$<br>$\frac{1}{4}HH ; \frac{1}{4}DH \text{ or } HD ; \frac{1}{16}DD$ |
| V For parents HD or DH   | $\frac{1}{4}DD ; \frac{1}{2}DH \text{ or } HD ; \frac{1}{4}HH$  |
| VI For parents DD        | DD  |

For each of these types we require a separate card pack with appropriately specified values of  $a_{0.4}$ ,  $b_{0.4}$ ,  $c_{0.4}$ , etc., above, e.g. for IV  $a_{0.4} = \frac{1}{16} = i_{0.4} = d_{0.4} = e_{0.4}$ ;  $b_{0.4} = c_{0.4} = \frac{1}{8} = g_{0.4} = h_{0.4}$ ; and  $f_{0.4} = \frac{1}{4}$ . We shall also need to express the relative frequencies of fraternities I-IV by  $f_i = f_1, f_2, \dots, f_6$ . These assign the probability of choosing at random a card pack of the appropriate type (*first stage*), before choosing a card (*second stage*) as a ticket of entry to the appropriate urn (A-F) from which (*third stage*) we extract a single ball.

*The Frequency Approach.* As stated, Mendel cites the result embodied in (v); but he reached it by reasoning at a much less formal level. He says in effect: let us put the case at its worst, starting with a generation in which there are *no* homozygous genotypes. Then all matings are HH and half the next generation must be heterozygotes. The rest will be homozygotes. So only half of the subsequent generation can have any heterozygous offspring and only half of such offspring will in fact be heterozygous. In this generation the proportion of heterozygotes will thus be  $\frac{1}{4}$ . If we denote by  $H_0 = 1$  the proportion of heterozygotes in the initial generation,  $H_1 = \frac{1}{2}$  in the next,  $H_2 = \frac{1}{4}$  in the next, and so on, we may therefore put  $H_n = 2^{-n}$  as the proportion of heterozygotes in the  $n$ th or less than one in a thousand after 10 generations of self-fertilisation.

There was no noteworthy advance of this aspect of theoretical genetics till the publication of a memoir by Jennings (1917), disclosing remarkable results, including genotypic series for brother-sister mating and simple assortative mating. Like Mendel, Jennings proceeds from an arbitrary origin of population structure and, also like Mendel, invokes no calculus more elaborate than common-sense arithmetic. His memoir is the foundation of later and more formal developments of the genetical theory of population by Haldane, Sewall Wright, Dahlberg, Hogben and others, all of whom explicitly enlist the classical theory of probability. That Jennings could travel so far without its help raises the question: in what sense is the formal calculus of probability essential and with what end in view do we usefully invoke classical models such as the foregoing. The answer to the second part of the question is unequivocal. We do so, more or less profitably, to steer our way

with as little effort as need be through a sequence of otherwise confusing operations; but the mere fact that we may then be talking about urns or card packs does not supply an answer to the first part.

To find one, let us now go back to the earliest class of problems which geneticists had to tackle. With no evidence to the contrary, we may assume that Mendel had not made a study of the algebraic theory of probability, his casual reference to *chance* being no more than current idiom. In any event, a corresponding statement would certainly be true of some of his foremost expositors among the pioneers of modern genetics, as the writer knew them in his student days. One has at least ample material for studying how they did in fact approach as teachers the class of problems subsumed in the epoch-making publication of the *Mechanism of Mendelian Inheritance* (1915) by Morgan, Muller, Sturtevant and Bridges; and we can retrace all their steps. We do not then start at the abstract level of Darwin's pangens. We proceed from a background of factual experience, more especially:

(a) that small samples of progeny may depart widely from the Mendelian ratios which provide a satisfactory description of a large pool of such samples;

(b) no visible characteristics of spermatozoa suggest that the genetic equipment of one class of sperms produced by a father has any influence on the relative frequency with which sperms so specified fertilise the egg;

(c) for all differences referable to autosomal gene substitutions the results of reciprocal crosses are identical.

We shall now advance the following postulates as a basis for a hypothesis which must stand or fall by the issue of experiment.

*Postulate 1.* In accordance with (a), we shall concede that statements we make are conceptually precise in the limit only;

*Postulate 2.* In accordance with (b), we shall assert that any two spermatozoa have the same opportunity to fertilise one and the same egg;

*Postulate 3.* We shall likewise assert that one and the same sperm has the same opportunity of fertilising any two ova to which it has simultaneous access.

# MENDELISM AND THE MEANING OF PROBABILITY

For heuristic purposes we may temporarily disregard our first postulate. We shall assume a set-up in which there are three eggs  $A_1$ ,  $A_2$ , and  $a_1$  available to five sperms  $A_3$ ,  $A_4$ ,  $A_5$ ,  $a_2$  and  $a_3$ . If we lay out all possibilities gridwise, our picture of what *may* happen in fertilisation takes shape as below:

	$A_3$	$A_4$	$A_5$	$a_2$	$a_3$
$A_1$	$A_1A_3$	$A_1A_4$	$A_1A_5$	$A_1a_2$	$A_1a_3$
$A_2$	$A_2A_3$	$A_2A_4$	$A_2A_5$	$A_2a_2$	$A_2a_3$
$a_1$	$a_1A_3$	$a_1A_4$	$a_1A_5$	$a_1a_2$	$a_1a_3$

If our only concern is with the class designation  $A$  and  $a$ , we may condense our grid of equal opportunity by assigning to each cell a score of unity as follows:

	$A$	$a$
$A$	$AA$	$Aa$
$a$	$aA$	$aa$

	$A$	$a$
$A$	6	4
$a$	3	2

We then see that we can represent the *proportionate possibilities* consistent with the equipartition of opportunity for association for the 15 different zygotes in the original table, as on the left below. On the right, we attach to each class a border score specifying its proportionate contribution to the appropriate total population (sperm or egg).

	$A$	$a$
$A$	$AA = \frac{6}{15}$	$Aa = \frac{4}{15}$
$a$	$aA = \frac{3}{15}$	$aa = \frac{2}{15}$

	$\frac{3}{5}$	$\frac{2}{5}$
$\frac{2}{3}$	$\frac{6}{15}$	$\frac{4}{15}$
$\frac{1}{3}$	$\frac{3}{15}$	$\frac{2}{15}$

We have now exhibited one of the two fundamental operations of our calculus. The other emerges, if we set out the result as below :

<i>AA</i>	<i>Aa or aA</i>	<i>aa</i>	<i>Total</i>
$\frac{6}{15}$	$\frac{7}{15}$	$\frac{2}{15}$	1

With due regard to Postulate 1, such, in the context of Mendel's monohybrid experiments, is the genesis of what we call the theorems of multiplication and addition in the calculus of probability; but we have reached our goal without using a word redolent with subjective associations. We have done so by recourse to postulates referable only to external events completely independent of our convictions and sentiments. If we care to designate as probabilities the proportionate possibilities defined as above, we must therefore renounce any disposition to locate probability so conceived in the mind. Whether the earliest expositors of genetical theory relied on the visual aid of the grid to exhibit the operations appropriate to the postulates because they were unfamiliar with the algebraic theory of probability, because they rightly assumed that most of their pupils would be, or because the word probability had become tarnished by the metaphysical doctrine of insufficient reason, are questions to which we need not seek a definite answer. What is more material to the choice of our theme is that the foregoing schema records how they actually did expound the theory of the gene. With appropriate emendations it contains all the necessary ingredients for a genetical theory of populations. It is therefore not surprising that the latter developed so far without explicit identification of its operations with those of the classical theory of risks in games of chance.

Without asking any more of our grid than it can do for us, we shall now examine another consequence of Mendel's hypothesis. With this end in view, we may usefully make explicit a postulate which we might otherwise regard as redundant. We shall say :

*Postulate 4.* For parents of given genotype the proportions of offspring of a specified genotype are the same for all birth ranks.



# MENDELISM AND THE MEANING OF PROBABILITY

For illustrative purposes, we may then apply our grid procedure to the situation in which the parental mating is RH, referable to a single autosomal gene substitution. In the long run, half the offspring will then be R and half will be H.

For 2-fold fraternities classified w.r.t. both genotype and birth rank, we may set out the table as follows:

		1st birth	
		R	H
2nd birth	R	$\frac{1}{4}$	$\frac{1}{4}$
	H	$\frac{1}{4}$	$\frac{1}{4}$
HH		RH. or HR.	
$\frac{1}{4}$		$\frac{1}{2}$	
		RR	
		$\frac{1}{4}$	

Alternatively, we may score the result in terms of numbers of recessives per 2-fold fraternity as:

<i>R-score</i>	0	1	2
<i>Frequency</i>	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Our fourth postulate implies that each result for the first two birth ranks will be equally often associated with either possibility for the third. We may set this out fully thus:

	HH	RH	HR	RR										
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$										
H $\frac{1}{2}$	HHH $\frac{1}{8}$	HRH $\frac{1}{8}$	HHR $\frac{1}{8}$	HRR $\frac{1}{8}$										
R $\frac{1}{2}$	RHH $\frac{1}{8}$	RRH $\frac{1}{8}$	RHR $\frac{1}{8}$	RRR $\frac{1}{8}$										
<table> <tr> <td><i>R-score</i></td> <td>0</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td><i>Frequency</i></td> <td><math>\frac{1}{8}</math></td> <td><math>\frac{3}{8}</math></td> <td><math>\frac{3}{8}</math></td> <td><math>\frac{1}{8}</math></td> </tr> </table>					<i>R-score</i>	0	1	2	3	<i>Frequency</i>	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
<i>R-score</i>	0	1	2	3										
<i>Frequency</i>	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$										

# STATISTICAL THEORY

More economically, we may now set out our grid for the result of a *fourth* birth with appropriate border R-scores as:

	0	1	2	3
0	0 $\frac{1}{16}$	1 $\frac{3}{16}$	2 $\frac{3}{16}$	3 $\frac{1}{16}$
1	1 $\frac{1}{16}$	2 $\frac{3}{16}$	3 $\frac{3}{16}$	4 $\frac{1}{16}$

Whence for the 4-fold fraternity we have:

<i>R-score</i>	0	1	2	3	4
<i>Frequency</i>	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

If we summarise the foregoing results, we have now the pattern for the *r*-fold fraternity:

	0	1	2	3	4
<i>1-fold fraternity</i>	$\frac{1}{2}$	$\frac{1}{2}$	..	..	..
<i>2-fold</i> „	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	..	..
<i>3-fold</i> „	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	..
<i>4-fold</i> „	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Thus for *r*-fold fraternities of matings RH, the long-run frequency of fraternities with 0, 1, 2 .... *r* recessives are successive terms of  $(\frac{1}{2} + \frac{1}{2})^r$ , i.e.:

0	1	2	3	.....
$\frac{1}{2^r}$	$\frac{r}{2^r}$	$\frac{r(r-1)}{2 \cdot 2^r}$	$\frac{r(r-1)(r-2)}{3 \cdot 2 \cdot 2^r}$	.....

For fraternities of 10 this yields:

0	1	2	3	4	5	6	7	8	9	10
$\frac{1}{1024}$	$\frac{10}{1024}$	$\frac{45}{1024}$	$\frac{120}{1024}$	$\frac{210}{1024}$	$\frac{252}{1024}$	$\frac{210}{1024}$	$\frac{120}{1024}$	$\frac{45}{1024}$	$\frac{10}{1024}$	$\frac{1}{1024}$

We now recall Postulate 1. We shall talk only about what happens when our numbers are very large, or *in the limit*, if we need to be fastidious. With that reservation, we shall say of a sufficiently large number of fraternities of ten with parents R.H: one out of every 1,024 of such fraternities will contain *no* recessive members. Let us then suppose we have before us a 10-fold fraternity with no recessive members. If we actually know that the parents are R.H. we also know that we are the privileged spectators of an event we have roughly a one in a thousand chance of witnessing. We may agreeably suppose that we now imbibe two or three tots of Scotch at the bar, lose our notes and no longer feel quite so sure that the parents were indeed R.H. Are we to dismiss the possibility that we may be the recipients of this privileged opportunity? The answer is that we cannot do so, unless we hypostatise our own ignorance as the pacemaker of the external situation. Otherwise, nothing which our postulates imply gives us the slightest justification for making any such stupendously irrelevant decision. Our hypothesis must stand or fall within the framework of its original terms of reference which include Postulate 1; and this excludes our title to make any statement about a *single* 10-fold fraternity in contradistinction to a denumerable infinitude of 10-fold fraternities, i.e. for practical purposes a very large number. We are then accepting the hypothesis on its own terms—as indeed did Mendel and his first exponents—by pooling the results of individual matings.

By itself, the outcome of such pooling does not suffice to justify the assertion that the hypothesis works. To say that a hypothesis works in the experimental sciences means more than asserting its consistency with known facts, especially if we rely on such facts to suggest the background model. We ask also that it shall lead us to new and unforeseen consequences which are verifiable. The theory of the gene has abundantly vindicated its claim to inclusion in the corpus of scientific knowledge by providing practical recipes for establishing new stocks of guaranteed purity by procedures previously unknown to practitioners of animal and plant breeding.

In the same way, we accept a physical hypothesis on its own terms whether the model enlisted in its construction is a

stochastic model or a non-stochastic model; and I have yet to learn that the physical chemist consults one of the new departments for the Design of Experiments in a University to help him to decide whether the Kinetic Theory gives or does not give a satisfactory account of the phenomena of freezing-point depression. If we seek a criterion for acceptance or rejection of any hypothesis in terms of the rarity of an event it can admittedly accommodate, we have indeed stepped into another domain, that of the Calculus of Judgments; and we have raised a new issue which invites enquiry on its own merits. Such will be the theme of our next chapter.

*The Two Probabilities.* A just assessment of the use of the theory of probability in the domain of the Calculus of Aggregates will clarify the task we shall then undertake. To get into focus the relevant issues which have emerged in this chapter, let us recall the argument of Chapter Two. We there distinguished between two ways of defining probability, the *formal* and the *empirical*. If we rely on experience to justify the choice of the model to which the formal definition is relevant, a preference for one or the other is equally consistent with a behaviourist outlook; and convenience for expository usage will dictate our choice. If we do not explicitly state the need for such verification of the relevance of the formal definition, the distinction becomes a challenge.

In the preface to the first volume of his treatise on the theory Carnap does indeed distinguish between different *probabilities*, one formal and ostensibly relevant to matters of judgment, one empirical and relevant to external events such as those which our present theme subsumes. The behaviourist will not dispute the contention that a distinction is necessary to describe two domains of verbal behaviour; and will even advance a plea for the substitution of separate vocables for Carnap's *probability one* and *probability two*; but the behaviourist will not be able to accept the compromise Carnap proposes. It is exceptionable, if only because theoretical statisticians (p. 22), are all too ready to invoke the successful application of a stochastic calculus in the domain of events to justify their claims to dictate rules of inference appropriate to the interpretation of experimental data. Thus the consumer naturally

concludes that such self-justification derives its sanction from one and the same usage of the term probability.

I have already cited Wilks to show how easily misunderstanding of this sort may arise; and I disclaim any intention of blame if I here cite another example. Referring explicitly to the contemporary controversy concerning the credentials of modern statistics with a bouquet *en passant* to Kolmogorov's "complete axiomatic treatment of the foundations of probability theory," Feller (1950),\* also a distinguished American mathematician, complains that "an *unfortunate publicity*" has been "given to discussions of the so-called foundations of probability and thus the erroneous impression was created that essential disagreement *can* exist among mathematicians" (*italics inserted*). He then assures us that "there exists no disagreement concerning mathematical facts" (pp. 6-7), an assertion which is not open to debate, if we choose to regard the theory of probability as a branch of pure mathematics. In that event, any axioms are admissible, if the outcome is self-consistent.

Such a view of its scope is difficult to square with the same author's claims for "the success of the modern mathematical theory" in the domain of practice, unless the criterion of success is the verifiability of the *end-product*. If so, the only domain in which the end manifestly justifies the means is the domain of Maxwell and Mendel. Contrariwise, the consumer should have the last word about whether experience justifies the *initial* axioms when the statistician claims the right to prescribe how to interpret and design his experiments. In that event, the consumer will be wise to examine such claims unmoved by the emotive appeal of Feller's plea (*op. cit.*, p. 7):

It is easy to decry theories as impractical. The foundations of practical things of today were so decried only yesterday, and the theories which will be practical tomorrow are branded as valueless abstract games by the practical men of today.

If it happens that the consumer has no disposition to rely on recipes of inference deriving no sanction from experience,

\* *Probability Theory and its Applications*, Vol. I.

it is his right both to know whether such recipes are or are not referable to a definition of probability consistent with his requirements and to know that recipes referable to a purely subjective definition of probability have no title to share the spoils of the hard-won victories of a calculus conceived in empirical terms. From one viewpoint, Carnap's contribution is therefore salutary. If it receives the publicity it deserves, its only effect will be to infect the investigator in the laboratory and in the field with a sceptical temper towards the claims of statistical theory, a temper more in keeping with the traditional claim that science is the last defence of intellectual freedom in its perennial conflict with arbitrary authority.

PART IV

---

*The Calculus of Judgments*





## CHAPTER FOURTEEN

### STATISTICAL PRUDENCE AND STATISTICAL INFERENCE

TOWARDS THE CLOSE of the last chapter we touched on a question at the core of contemporary controversy in statistical theory: what bearing, if any, has the rarity of an observable occurrence as prescribed by an appropriate stochastic hypothesis on our legitimate grounds for accepting or rejecting the latter when we have already witnessed the former? The form of the answer we deem to be appropriate will define what we here conceive to be the proper terms of reference of a Calculus of Judgments, i.e. *statistical inference* as some contemporary writers use the term. Such is the theme of this chapter and of those that follow.

At the outset, it will forestall misunderstanding if we concede that some contemporary writers use the term statistical inference in a wider sense than as defined above, embracing any sort of reasoning which takes within its scope considerations referable to a calculus of probability. Such usage is regrettable, because the form of words suggests a more radical difference. By inference in the traditional sense of the term as used by the logicians of science from Bacon to Mill and Jevons, we signify rules which *unreservedly* lead to correct conclusions. By statistical inference, or, as we shall later say, *stochastic induction*, we here imply rules of reasoning which lead to correct conclusions subject to a reservation expressible in stochastic terms. We relinquish the claim that any such rule *always* leads to a correct conclusion. All we claim for such a rule is that it guarantees an assignable probability of correct or false assertion, if we apply it consistently to the relevant class of situations.

By an assignable probability we then mean the proportion of true or false assertions to which the rule will lead us in an endless sequence of such situations, and the theory of probability can endorse such a rule if, and only if, the assemblage of data for our observational record is the outcome of randomwise

sampling. For the time being, I shall assume that this is so, and shall reserve till Chapter 19 an examination of what obligations the assumption imposes. On that understanding, we may specify the proportion of false assertions as the *uncertainty safeguard* ( $P_f$ ) of the rule with an upper acceptable limit  $P_f \leq \alpha$ . Alternatively, we may assign a lower acceptable limit  $P_t = (1 - P_f) \geq (1 - \alpha)$  to the proportion of true statements which consistent application of the rule will endorse in the long run. We may then speak of such a lower limit as the *stochastic credibility* of the rule. What limit we deem to be acceptable in this context is not a logical issue. If we commonly set  $\alpha = 0.05$  for illustrative purposes in the course of the chapters which follow, we do so because it is a widely current convention which will help some readers to feel at home in familiar territory.

In pursuing our theme on this understanding and historically as heretofore, we may conveniently distinguish between two sorts of statistical inference as *test procedure* and *interval estimation*. The dichotomy is provisional. For we shall later see that interval estimation embraces test procedure as a special case. Meanwhile, two difficulties beset our task. One arises from the circumstance that a strictly behaviourist formulation of the terms of reference of a stochastic calculus of judgments, as in the preceding paragraph, is wholly consistent with the attitude to test procedure adopted by only one of two different schools of doctrine; and its implications have come clearly into focus only since the contemporary controversy touched on in Chapter 1 has forced the contestants to make explicit latent assumptions lazily embraced by their predecessors. Contemporaneously, claims put forward by exponents of the alternative school have changed in response to criticism not previously anticipated. Thus our foothold on fact is insecure when we seek to evaluate the original intentions of the authors with due regard to the way in which the consumer has meanwhile interpreted them.

Since emphasis on differences of test prescription and on what each type of test can accomplish has shifted in the turmoil of discussion, it is not easy to circumscribe acceptably the connotation of the term *decision test* as defined by Wald on the

basis of views first advanced by J. Neyman and E. S. Pearson in contradistinction to the term *significance test* as prescribed by the school of R. A. Fisher in the older tradition of Karl Pearson and Udny Yule. In this chapter, we shall therefore disclaim any attempt to verbalise the distinction with finality. Instead, we shall restrict our attention to the historical background of the controversy. Against this background, we may formulate a preliminary statement of what a significance test is and of what are its claims to usefulness, deferring a definition of the decision test as such to Chapter 15 in the context of an examination of appropriate model situations. Comments here stated *en passant* will therefore anticipate formal reasoning advanced more fully in what follows. Accordingly the reader may find it advantageous to defer careful reading of this chapter to a later stage.

We may trace the beginnings of a view of test procedure widely current during the twenties and the thirties of this century to the practice of citing with point-estimates in surveying and in the physical sciences an *empirical* assessment *either* of the precision index ( $h$ ) prescribed by the Gaussian, as we now say *normal*, law of error *or* of some parameter of the distribution related thereto, more especially the so-called *probable error*. For a normal error distribution of known precision (p. 164), the probable error ( $p.e.$ ) defines a continuous class of corresponding negative and positive deviations from the mean bounding half its total area. In sampling randomwise, equal probability thus attaches to the occurrence of deviations numerically less than and deviations numerically greater than the  $p.e.$  The probability that a deviation from the mean will numerically exceed no more than three times the  $p.e.$  is about 0.95.

Throughout the nineteenth century, those who followed this procedure did so with little disposition to claim that it embodies a major innovation of scientific reasoning. Nor did they need to do so. In the domain of precision instruments and of mechanical defects, it would be easy to cite a variety of situations in which the specification of a deviation from the putative true value of a dimension or constant has a utility none the less admissible because no professional logician would endorse its

title to disclose a new formula for scientific induction. A single fictitious and over-simplified example from the contemporary setting of quality control (statistical inspection) at its most elementary ( $p$ -chart) level will suffice to clarify what we can legitimately and usefully say about a probable error in the Gaussian domain.

We shall suppose that: (a) a machine turns out lengths of metal wire of thickness guaranteed to lie within 0.98 and 1.02 mm.; (b) the variation of the thickness under normal working conditions is approximately normal with a probable error 0.004 mm. about mean 1.0 mm. Thus values which numerically exceed the mean by more than the guaranteed 0.02 mm. will turn up in only about one four hundredth of a very long sequence of samples during the normal production process. So long as the machine delivers the guaranteed product there is no need to interfere with it; but it may well run off a length of wire 1.025 mm. thick. The production engineer has then to face a dilemma. In which of two ways is he to interpret the event? He will know that the mishap may well be an extreme example of the uncontrollable vagaries of the mechanical set-up. He may also legitimately suspect that the machine has developed a fault which calls for repair. Without incurring the risk of having a large consignment of defective products on his hands, he can immediately settle the issue by an overhaul; but this spells needless loss of time if his suspicion is groundless. A plausible rationalisation of his choice therefore presupposes the possibility of balancing the cost of taking a wise precaution which may be unnecessary against the possible penalty of failing to do so.

His predicament is indeed on all fours with that of the reflective householder whose dog does not habitually bark in the night. There may be no burglar, when the dog does indeed bark. On the other hand, the risk may be worth the effort of getting out of bed, if there is enough portable property of value to conserve. Neither the householder who justifies his decision to make a search on these terms, i.e. as a wise rule of conduct, nor the engineer, who justifies the procedure of overhauling the machine as the less calamitous of two acts of choice, necessarily commits himself to an *inference* in the sense in which

logicians traditionally use the term. To be sure, we might say of the householder and of the production engineer that each is testing the truth of alternative hypotheses; but the recognition of the rarity of the relevant event as a danger signal rather than the interpretation of its outcome dictates the choice of the test procedure; and the finality of the selected test has nothing to do with the rarity of the event which prompts its choice. If the dog's bark evokes the decision of the householder to go downstairs, he will admittedly be able to infer the truth or falsity of the hypothesis that a burglar is on the premises; but if he decides to remain in bed, he will also know before breakfast whether there has been an entry.

It suffices to speak of the engineer's rule as a rule of *statistical prudence*; but we may reasonably conclude that the theoretical statistician means something more than this when he uses a form of words so portentous as *statistical inference*. Indeed, some contemporary writers pinpoint such a distinction by using the term *conditional*, in contradistinction to unconditional, inference for what we here refer to as statistical prudence. Contrariwise, others make no such distinction and seemingly claim for what a test procedure can accomplish little more than the prescription of a wholesome discipline. If so, the research worker who embraces a test procedure presumably undertakes what may be a time-consuming programme of laborious computations in the mistaken belief that the statistician has much more to offer.

When we speak of testing a hypothesis, the form of words suggests a procedure for endorsing a valid verdict for or against its title to subsume new information worthy to take its place in the enduring corpus of scientific knowledge. The consumer has therefore the right to know that he is doing nothing of the sort, if the test is merely a *screening* convention to check rash decisions or to arbitrate on the advisability of following up a plausible hunch. An experienced investigator, with no illusions about the practicability of formulating risks relevant to further effort in numerically intelligible terms consistent with the professional ethic of scientific research, may accordingly prefer to rely on common sense, if statistical theory has nothing better to confer. The truth is that the distinction between statistical inference as an interpretative device defined at the beginning

of this chapter and statistical prudence conceived as a discipline to forestall rash judgments did not become clear till Fisher repudiated the scholium of Bayes in his *Mathematical Foundations of Theoretical Statistics* (1921). This evoked a vigorous rejoinder from Jeffreys, whose confessedly idealistic approach to the theory of probability would not otherwise fall within the scope of an assessment of the present crisis from a behaviourist viewpoint.

Jeffreys propounded what Anscombe (*op. cit.*, p. 24) calls a *proper theory of induction*, i.e. a system of which the terms of reference are at least intelligible to those who find the initial assumptions acceptable. In controversy with him, Fisher consistently refused to commit himself explicitly to a definition of probability located in the mind, and, by so doing, enlisted the sympathy of contemporaries to whom such a formulation would be repugnant; and nothing in his earlier publications unequivocally advances the claims of test procedure as an innovation of logical technique. As controversy over the scholium sharpened, we can trace the emergence of a new *motif* in his writings. In his *Mathematical Foundations* referred to above, he elaborates the formal algebra of a test procedure essentially *en rapport* with that of his predecessors in the tradition of Karl Pearson, and in an idiom which does not provoke the professional logician to assent or denial. Only when controversy with Jeffreys had clarified some of the cruder implications of abandoning the doctrine of Laplace did his assertion of the claims of *uncertain inference* explicitly annex a procedure which the logician might reasonably have regarded as the more fitting preserve of a disciplinary precaution.

The treatise of Jules Gavarret mentioned elsewhere, signals what seems to be the earliest intrusion of the probable error into the domain of natural variation. Therein Gavarret propounds a statistical approach to the evaluation of the efficacy of a treatment. It is in essence the significance test as commonly conceived in the twenties and thirties of our own century; but it seems to have exercised little influence on medical research in the author's own lifetime, perhaps because overshadowed by spectacular contemporaneous advances in experimental physiology. Meanwhile, there was a soil more

favourable than medicine for the seed of the word. In the domain of natural variation, comparative anatomy was in the ascendant. Under the impact of the Darwinian doctrine in the setting of Galton's racialist creed and of the controversy over slavery in America, anthropometry became a fashionable academic playground. Thus we find one of the earliest examples of the use of the terms *test* and *significant* in their now current meaning in a memoir by J. Venn (1888) in the *Journal of the Anthropological Institute* (pp. 147-8):

But something more than this must be attempted. When we are dealing with statistics, we ought to be able not merely to say vaguely that the difference does or does not seem significant to us, but we ought to have some test as to what difference would be significant. For this purpose appeal must be made to the theory of Probability. Suppose that we have a large number of measures of any kind, grouping themselves about their mean in the way familiar to every statistician, their degree of dispersion about this mean being assigned by the determination of their "probable error." . . . For instance, the difference in the mean length of clear vision between the A's and the C's is about an inch and a quarter; that between the same classes, of the age of 24, is slightly more, viz. about an inch and one-third. But the former is the difference between the means of 258 and 361, the latter that between means of 25 and 13. By the formula above given we find that the respective probable errors of the differences between these means are one-twelfth and one-third of 3.7 inches, i.e. about 0.3 inches and 1.2 inches. The latter is almost exactly the observed difference, which is therefore seen to be quite insignificant. The former is about one-quarter of the observed difference, which is therefore highly significant; for the odds are about 25 to 1 that a measure of any kind shall not deviate by three times its probable error.

The above remarks are somewhat technical, but their gist is readily comprehensible. They inform us which of the differences in the above tables are permanent and significant, in the sense that we may be tolerably confident that if we took another similar batch we should find a similar difference; and which of them are merely transient and insignificant, in the sense that another similar batch is about as likely as not to reverse the conclusion we have obtained.

All this adds up to little in terms of the world's work, because anthropometry had then (as now) scanty, if any, practical value

other than for manufacturers of ready-made wearing apparel or of school furniture. That reliance on test procedures so conceived suddenly becomes so universally *de rigueur* among experimental biologists in the twenties and thirties of our own century is an enigma which admits of no wholly satisfactory solution; but one clue to the mystery may be the fact that the influence of men such as Farr had already popularised interest in descriptive medical statistics in anticipation of public legislation to enforce immunisation procedures. By the end of the nineteenth century, the vaccination issue had indeed become the focus of a vehement public debate which provided a propitious setting for the reception of the views Gavarret had earlier expressed. Had it been true that men of science were themselves of one accord in the polemics of the last decade of the nineteenth and of the first decade of the twentieth century, the outcome might have been otherwise; but the opponents of vaccination could themselves claim enthusiastic supporters among biologists, for instance Alfred Russel Wallace. In such circumstances, a counsel of moderation could prevail against reckless washing of dirty linen in public only by taking the issue to a higher court of appeal with a better prospect of assembling a unanimous bench of respected judges.

Such was the situation when Yule and Greenwood published a memoir which was the curtain-raiser to the appearance on the stage of the statistician as arbitrator on matters about which trained observers disagree. The publication (*Proc. Roy. Soc. Med.*, Vol. VIII) of the last named authors bore the title *The Statistics of Anti-typhoid and Anti-cholera Inoculations and the Interpretation of such Statistics in general*. It appeared (1915) in the middle of a world catastrophe, at the end of which the notion of significance becomes for the first time prominent in a new *genre* of statistical textbooks, such as those of Bowley and of Caradog Jones. Meanwhile a generation suckled on the *Grammar of Science* and attuned to the controversies of *Biometrika* were rising to positions of influence in the biological world. Of such were Raymond Pearl in America and in England Major Greenwood himself.

By then, food shortage at the end of the First World War had catalysed interest both in deficiency diseases and in agri-



cultural output. The former merits comment because research on the vitamins enlisted methods of bio-assay which rely on group averages in contradistinction to one-to-one correspondence of stimulus and response of one and the same individual. Food production is especially relevant because there was now a ready audience for the honest broker of the vaccination controversy, when R. A. Fisher (1925), then engrossed in fertiliser records, published a well-known textbook ostensibly addressed to research workers in general, though in fact almost exclusively concerned with the agricultural field trial. Almost overnight it became a best seller. In America, agricultural statisticians, there led by Snedecor, became the most enthusiastic converts to an evangel which is indeed a still unanswered challenge to Bernard's teaching and to the Baconian recipe.

*Pari passu* throughout the thirties, the use of statistical tests in the experimental sciences became more fashionable; but there was little inclination to probe their several claims in the domains of error and of natural variation. Meanwhile medicine remained aloof from the statistical approach to the therapeutic or to the prophylactic trial until unprecedented production of new synthetic drugs and of antibiotics on the eve of, and during, a second world war. An informed public now eagerly awaited a verdict on their merits. By that time, controversy concerning the current statistical recipes had become vocal and vigorous in mathematical circles; but its echoes did not penetrate the pathological laboratory. It would be unnecessary to write this book, if many men of science did as yet realise how little the College of Cardinals can agree about the rubric.

I have deliberately used the term *field trial* in the foregoing remarks. Such at first, and rightly, was the designation of what the statistician of a later vintage refers to as *experiment*; but the acceptability of the new title is instructive, because it focuses attention on a confusion of aims. The twenties and thirties witnessed rapid advances in techniques of bio-assay first as tools of pure research in dietetics and in endocrinology, later as the means of accrediting the reliability of products in commercial production. Juxtaposition of statistical tests employed in comparable circumstances in the milieu of the new pharmacy seemingly encouraged the belief that there is a common

denominator for the objectives which a commercial corporation and a disinterested scientific worker pursue. As Abraham Wald explicitly defines it, and as several expositors of Fisher's methods for the use of pharmacologists define it by implication, statistical inference is this common denominator.

Perusal of the earliest prescriptions for the testing of hypotheses may well leave the reader in considerable doubt about what the pioneers did claim. What is clear is that they carried over from the proper domain of the Gaussian theory of point-estimation a body of concepts with disputable relevance to a new class of situations, when the immunisation controversy recruited the statistician as the arbitrator of truth and falsehood. The ill-fated marriage of the Gaussian theory of error with the empirical study of populations had already generated a sturdy progeny of misconceptions dealt with elsewhere. Hence we may record the event without bewilderment, even if the *ipsissima verba* of Karl Pearson, who played a leading role in the background of the debate, do not greatly help us to formulate a clear distinction between the scope of scientific induction as interpreted in the English tradition from Bacon to J. S. Mill and induction restated in terms of stochastic theory. We encounter an early use of the now familiar expression *significant difference* in a discussion (*Biometrika*, Vol. 8, 1911, p. 250) on the *Probability that two Independent Distributions of Frequency are really Samples from the same Population*; and it would be difficult to select a more forceful illustration of the *Backward Look* than in the following citation therefrom:

In a memoir contributed to the *Phil. Mag.*, 1900 (Vol. 50, p. 157), I have dealt with the problem of the probability that a given distribution of frequency was a sample from a *known* population. That investigation was the basis of my treatment of the "goodness of fit" of theory and frequency samples. The present problem is of a somewhat different kind, but is essentially as important in character. We have two samples, and *a priori* they may be of the same population or of different populations; we desire to find out what is the probability that they are random samples of the same population. This population is one, however, of which we have no *a priori* experience. It is quite easy to state innumerable problems in which such knowledge is desirable. We

have two records of the number of rooms in houses where (i) a case of cancer has occurred, (ii) a case of tuberculosis has occurred; the number of cases of each disease may be quite different, and we may not be acquainted with the frequency distribution of the number of rooms in the given district. *What is the chance that there is a significant difference in the tuberculosis and the cancer houses?* Or again, we have a frequency distribution of the interval in days between bite and onset of rabies in two populations of bitten persons (i) who have been and (ii) who have not been inoculated in the interval. What is the probability that the inoculation has modified the interval? Many other illustrations will occur to those who are dealing with statistics, but the above will suffice to indicate the nature of the problems I have in view. (*Italics inserted.*)

For Pearson, as for Venn (p. 61), a difference is *significant*, in the sense that it casts doubt on a hypothesis, if the chance of its occurrence is very small on the assumption that the hypothesis is true. Thus it would not be easy to confuse so many issues with so few words as those of the query: "what is the chance that there is a significant difference in the tuberculosis and the cancer houses?" The publication of Yule and Greenwood (1915)\* cited above, defines significance in the same way. At the outset, the authors give their own interpretation of the question: "is there a significant difference between the attack or fatality rates of the two classes?" To them—as to a subsequent generation of consumers—this is strictly equivalent to asking: "is the observed difference greater than we could fairly attribute to the action of chance?" The answer they seek, but without attempting a definition of *fairly* in this context, is indeed an answer to a different question: how often "errors of sampling would lead to as great a discrepancy as or a greater discrepancy than that actually observed between *theory* and observation." (*Italics inserted.*)

One implication of the foregoing is that we can first look at

\* The first edition of Yule's *Theory of Statistics* published in 1911 devotes no more than three pages to illustrations of a method of testing the significance of a difference, i.e. the truth of the hypothesis that the universes from which different samples come are in all relevant particulars alike. The rest of the book deals with summarising indices devised in accordance with stochastic considerations in the manner of Quetelet with no explicit recognition of the need to distinguish between a rule of decision and a law of nature.

our sample and then decide what rule to apply. This we have already seen (p. 39, *et seq.*) to be wholly inconsistent with a behaviourist approach. Remarks on p. 171 have anticipated another objection, which we can get into focus more readily if we examine test prescription through the spectacles of the Forward Look. In effect, we then say that we shall reject a particular hypothesis if the deviation ( $X$ ) of the sample score from the mean ( $M_x$ ) of the distribution *numerically* exceeds a certain value  $X_r$ , so chosen that the probability assigned by the same hypothesis to sample scores in the range  $M_x \pm X_r$  is  $P_t = (1 - P_f) = (1 - \alpha)$ . The choice of a *score rejection criterion*, so defined, raises a debatable issue at the outset. If the rarity of an observed occurrence as prescribed by a particular hypothesis can indeed endorse any intelligible grounds for rejecting it, there is still no obvious reason, other than the culture-lag of the Quetelet mystique, to compel us to define a score rejection criterion in this *two-sided* way. If our real concern is with the possibility that inoculation may lower the attack rate, our criterion of rarity will be relevant to the end in view only if we define it in a *one-sided* way, i.e. in terms of score values less than one which we may denote alternatively as  $x$ , if the origin of the distribution is the least value (here zero)  $x$  may have, or  $-X_r$  if we transfer the origin to the mean for algebraic convenience.

The distinction is worthy of emphasis, because writers in the Yule-Fisher tradition, though commonly disposed to adopt the *modular* (two-sided) rejection criterion when the sampling distribution is symmetrical, otherwise employ a *vector* (one-sided) rejection criterion without explicitly disclosing the relevance of the algebraic properties of a particular sampling distribution to the factual content of the decision involved. The use of the word *theory* in the foregoing citation from Yule and Greenwood helps us to trace this confusion of aim to its source. What they here *signify* by theory is the long-run mean value of the sample difference, i.e. *zero* if the universes from which two samples come are alike w.r.t. all relevant particulars; but in this context the word is an unwarranted intruder from the proper domain of precision instruments. We cannot here assume that any universe parameter has a *true* value which would suffice to define the sample structure if we did not make mistakes.

The situation to which Yule and Greenwood refer is in no relevant respect comparable to the class of problems we discuss in the Gaussian domain. No parameter of the distribution has any special claim to dictate how to delimit from the class of all samples a particular sub-class to which we may be able to assign a very low frequency. Indeed, we shall see that the repudiation of such an arbitrary choice is an essential feature of the theory of test procedure expounded by E. S. Pearson, Neyman and Wald.

It goes without saying that the adverb *fairly*, as used by Yule and Greenwood, begs the whole question at issue. The algebra which the test invokes tells us how often we should encounter a certain and arbitrarily defined range of observed values on the conditional assumption that the samples come from like universes. Thus the mere fact that a particular observation is one of an arbitrarily delimited class of values, themselves collectively rare, is *ipso facto* consistent with the possibility that the hypothesis is true. In short, the so-called laws of chance provide for the very contingency which Yule and Greenwood invite us to regard as inconsistent with the truth of the test hypothesis. This being so, one may find it difficult to sympathise with Greenwood's dignified expression of grief in response to the ensuing and pertinent comment of his doughty clinical opponent Sir Almroth Wright:

It cannot be too clearly understood that the mathematical statistician has no such secret wells of wisdom to draw from, and that his science does not justify his going one step beyond the purely numerical statement that—as computed by him from the data he has selected as suitable for his purposes—the probabilities in favour of a particular difference being or not being due to the operation of chance are such and such. There need, therefore, be no hesitation in saying that when the mathematical statistician makes free with the terms *significant* and *non-significant*, he is simply taking upon himself a function to which he can lay no claim in his capacity as a mathematician.

It is thus equally impossible to extract any intelligible definition of test procedure from Karl Pearson's earlier pronouncements or to pin down the expressed views of Yule and Greenwood in the publication last cited to a decisive statement

which would confer on the term statistical inference any intention not implicit in the suggested alternative statistical prudence. The same is true of the book (see *Appendix IV*, p. 487, *et seq.*) which introduced statistical test procedure to a wider audience of investigators. In *Statistical Methods for Research Workers* (1925), destined to be the parent of a large fraternity of manuals setting forth the same techniques with exemplary material for the benefit of readers willing—and as it transpired, only too anxious—to take them on trust, Fisher's formulation of the rationale of the significance test neither discloses a new outlook explicitly nor clarifies views expressed by his predecessors. All that is novel is a refinement of the algebraic theory of the sampling distributions—with one notable exception embraced by Pearson's (1895) system of moment-fitting curves.

Without reading into the words of R. A. Fisher or of the authors cited above more than they would have conceded on second thoughts, we may none the less fairly define in broad outline under three headings the essential features of the *significance*, in contradistinction to the *decision*, test:

(i) we set up a single, the so-called null, hypothesis that one or more samples come from a hypothetical infinite population whose random sampling distribution is specifiable;

(ii) we decide to *reject* the hypothesis whenever the deviation of the sample score ( $X$ ) from the *mean* of the distribution so prescribed is such that  $P_f = \alpha$ , if  $P_f$  is the probability of meeting a score deviation numerically\* equal to or greater than  $X$ .

(iii) our reliance on the foregoing procedure places on us *no* obligation to specify in advance the size ( $r$ ) of the samples to which we propose to apply the test, and hence no pre-assigned rejection score criterion  $X_r$ , consistent with the agreed value of  $\alpha$ .

\* We shall return to this issue in Chapter Fifteen (p. 350 *et seq.*). As stated on p. 330, exponents of the test procedure are not wholly consistent w.r.t. the use of a 2-sided criterion here implicit in the word *numerically* in accordance with the choice of the mean as origin; but they give no factual reasons for adopting a one-sided criterion, when they do so.

For relevant source material with reference to (i) and (ii) the reader may consult *Appendix IV on Significance as Interpreted by the School of R. A. Fisher*. The third heading which draws attention to the most essential difference between the terms of reference of a significance, in contradistinction to a decision, test will occupy our attention in the ensuing chapter. There it will be necessary to examine the definition of the rejection criterion specified by (ii) in a more formal way than in what follows. Here our concern will be the content of (i) and (ii) in so far as they involve a still largely inarticulate conflict of interest intensified by amendments explicitly adopted to forestall criticism of the theoretical foundations of the test procedure without reconsideration of the saleability of the final product to a consumer fully acquainted with what the producer can now guarantee.

From this viewpoint, an issue raised in (ii) claims prior attention. We have assumed that the test merely prescribes when to *reject* the null hypothesis. Actually, the earliest exponents of the significance concept are by no means definite about this; and what seems to be the first wholly unequivocal pronouncement of R. A. Fisher (p. 500) is in the *Design of Experiments*. This first appeared after publications which prescribe another view of test procedure had already raised the question: when and in what sense can we legitimately accept the alternative to the null hypothesis? In the context referred to, Fisher's statement that the test outcome can sometimes disprove but never prove the truth of the null hypothesis is, to say the least, obscure; and it evades the sixty-four dollar question: what do we mean by proof in the domain of statistical inference? The reason for the afterthought will be clear enough at a later stage, when we shall examine (Chapter 15) the concept of test power.

If we do explicitly limit the terms of reference of our test procedure to rejection, it seems to me that we may interpret it in terms consistent with a behaviourist viewpoint in either of two ways:

- (a) in the phraseology of p. 346, we shall assign an uncertainty safeguard to a rule which is not *comprehensive*. In effect we say: in some situations we shall reserve judgment and

in others we shall make a decisive statement in accordance with a prescribed rule. We can then assign an uncertainty safeguard ( $P_f \leq \alpha$ ) which defines the probability of erroneous decision in a restricted class of situations; but we do so with no means of knowing how often the test will fail in the sense that we arrive at no decision at all.

(b) we shall impose on ourselves the self-denying ordinance of relinquishing the prescribed hypothesis only in exceptional circumstances as a disciplinary precaution against too ready acceptance.

Only the first of the two views here stated merits to rank as a technique of statistical inference in contradistinction to what we might more appropriately designate statistical prudence. Either way, the utility claimed for the performance of the test raises the question: have we any more reason for deciding when it is important to reject than for deciding when it is important to accept a particular hypothesis? This at least is an issue on which the consumer may rightly claim to voice an opinion; and it is scarcely deniable that the laboratory worker who invokes a test procedure conceived in terms of a unique null hypothesis commonly assumes that the test procedure justifies acceptance or rejection on equal terms. Indeed, the present writer can see no reason why the investigator should shoulder the responsibility of performing an elaborate drill of statistical computations, unless fortified by this belief. A few citations (*italics inserted*) from the good books which have taught the laboratory worker to do so will show that he has ample encouragement for his faith:

(i) Many scientific investigations involve the employment of the method of framing working hypotheses and testing them experimentally. As long as the experiments fail to disprove them, so long are the hypotheses accepted. This is the general method by which statistical inferences are made . . . the probability level of the observed difference is calculated accordingly. . . . The hypothesis is *accepted* if the level is fairly high and . . . if the level is low (say below 0.05) the hypothesis is rejected. (Tippett, *The Methods of Statistics*, 1931, pp. 69-70.)

(ii) In presenting the results of any test of significance the probability itself should be given. The reader is then in a position



to form his own opinion as to the justification of the *acceptance or rejection* of the hypothesis in question. (Mather, *Statistical Analysis in Biology*, 1942, p. 21.)

The last source is of special interest, since the book carries the *nihil obstat et imprimatur* of a Foreword by R. A. Fisher; and it is therefore pertinent to quote (2nd ed., 1946, p. 194) the author in a context\* which exhibits the investigator in the act of interpretation:

(iii) . . . which for 1 degree of freedom has a probability of 0.30-0.20, *showing that there is no interaction* between the classifications, i.e. that the type of water does *not* affect germination.

Snedecor, the most widely read exponent of the Fisher test battery, is less explicit, but does not dispel the belief that acceptance of something is the presumptive alternative to rejection of the null hypothesis:

(iv) Statistical evidence is not proof. Even after extensive sampling the investigator may not reject the null hypothesis when in fact the hypothesis is false. For example,  $m$  may not be zero in the population, yet natural variation may be great enough to confine  $t$  to a moderate value in any practicable size of sample. Therefore when one fails to reject the hypothesis he does not thereby conclude that  $m$  is zero. He decides only that  $m$  is *so small as to be unimportant to his investigation*. (Snedecor, *Statistical Methods*, 4th edn., 1946, p. 47.)

In Chapter 15 we shall see why we can legitimately draw no such conclusion as the one last stated without due regard to what Neyman calls the *power* of the test. In that event, what we deem to be small depends more on the size of the sample than on its importance to the investigator. The writer disowns any intention of carping criticism if seemingly implicit in citing twice from Mather's book. Mather writes less as a logician interested in the test credentials than as an investigator concerned with its utility. His attitude is of interest in this context mainly because his words make explicit the only terms in which a hard-headed investigator will presumably welcome a significance test procedure as an instrument for validifying

\* Here the null hypothesis is that there is no interaction.

conclusions reached in the laboratory or in the field. Should the laboratory worker turn to treatises written from the viewpoint of the mathematician, he will not necessarily find an interpretation of the use of the test inconsistent with the understanding that failure to reject is equivalent to acceptance:

(v) We begin by asserting that the hypothesis to be tested is true. . . . We may calculate the probability  $P(D > D_0)$  that the deviation  $D$  will exceed any given quantity  $D_0$ . . . . Let us choose  $P(D > D_0) = \epsilon$  where  $\epsilon$  is so small that we are prepared to regard it as *practically certain that an event of probability  $\epsilon$  will not occur in one single trial*. . . . If we find a value  $D > D_0$  this means that an event of probability  $\epsilon$  has presented itself. However, on our hypothesis such an event ought to be practically impossible in one single trial, and thus we must come to the conclusion that in this case our hypothesis has been *disproved by experience*. On the other hand, if we find a value  $D \leq D_0$  we shall be willing to *accept the hypothesis* as a reasonable interpretation of our data. . . . (Cramer, *Mathematical Methods of Statistics*, 1946.)

(vi) Improbable arrangements give clues to assignable causes; and excess of runs points to intentional mixing, a paucity of runs to intentional clustering. It is true that these conclusions are never foolproof. Even with perfect randomness improbable situations occur and may mislead us into a search for assignable causes. However this will be a rarity, and with an appropriate criterion we shall in actual practice be misled once in 100 times and *find* assignable causes 99 out of 100 times. (Feller, *An Introduction to Probability Theory and its Applications*. 1950.)

\* \* \* \*

These citations suffice to show that expositors of the significance test give the laboratory or field worker enough encouragement for overlooking the reservation specified in (ii) above (p. 332). That the investigator would indeed less readily embrace the type of test procedure they expound if fully aware of its implications will be clear enough if we now examine (i) against the background of the situation discussed by Yule and Greenwood in the publication already cited. The null hypothesis ( $H_0$ ) which circumscribes the prescription of the test procedure in accordance with (i) on p. 332 is that our two groups (treated and untreated) come from the *same*

*infinite hypothetical population.* This is the negation of the assertion that prophylactic inoculation is efficacious; but the main preoccupation of the investigator, who will commonly approach the task of carrying out the trial with a good hunch about the outcome, a hunch derived from preliminary experiments on related animals, from laboratory culture of bacteria or of viruses or from clinical observation, is in practice to establish the affirmative, i.e. to vindicate the credentials of a new instrument of preventive medicine. For what reason then should he or she be eager to take advantage of a test which can merely assign a low probability to erroneously asserting that the treatment is useless, but with no guarantee that the most likely result of applying it will be an open verdict, i.e. no verdict at all? We can justify the choice of our null hypothesis on such terms only from the disciplinary viewpoint defined by (b) on p. 334; but we are then using the idiom of statistical prudence rather than that of statistical inference. As Keynes (*op. cit.*, p. 300) remarks, the assumption

that it is a positive advantage to approach statistical evidence *without* preconceptions based on general grounds, because the temptation to "cook" the evidence will prove otherwise to be irresistible, has no *logical* basis and need only be considered when the partiality of an investigator is in doubt.

All procedures of the type under discussion, including the entire test battery of *Analysis of Variance* and *Analysis of Covariance* elaborated by R. A. Fisher and by his pupils, are referable in the last resort to the hypothesis that samples come from one and the same specified population in accordance with (i) of p. 332. If we ask why, the only reason offered is the reason given by Fisher in the context referred to above. So specified, the null hypothesis is seemingly unambiguous and on that account its algebraic formulation is tractable; but this butters no parsnips from the viewpoint of the laboratory worker who understands what he is doing. In an oblique reply (p. 498) to criticism of the significance test in these terms, Fisher admittedly invites the research worker to be the arbiter of the choice of the null hypothesis; but the motives which have promoted his own exploration of sampling distributions and the models set

forth by all his expositors are inconsistent with such freedom and with his own unequivocal expression of faith in the infinite hypothetical population as the keystone of the test theory edifice.

In the set-up of the clinical trial, choice of a null hypothesis appropriate to the operational intent is never consistent with (i), since it implies that the two samples come from different populations; and the only hypothesis consistent with (i) is that the mean ( $M_d$ ) of the sample score difference ( $d_m$ ) in randomwise extraction is zero ( $M_d = 0$ ). The intelligible alternative to our prescribed null hypothesis will be unambiguous, only if our main concern is to make a terminal statement of the form  $M_d \geq k$ , i.e. to the effect that prophylactic treatment lowers the proportion who succumb to attack by a target figure ( $k$  per cent) deemed sufficient to justify its adoption. As we shall later see, we must then define our uncertainty safeguard in the limiting form  $P_f \leq \alpha$  in contradistinction to the form specified in (ii) on p. 332; and this is true of any test prescription referable to a *discrete* sampling distribution, if we claim the right to prescribe a preassigned rejection criterion at any acceptable level.

The intelligible alternative ( $M_d \geq k$ ) to the null hypothesis that both treatment groups are samples from a single infinite hypothetical situation, signifies that each is a sample from a unique population; but Fisher's explicit statements proscribe the choice of a null hypothesis conceived in such terms. What thus emerges from the pivotal role of the infinite hypothetical population in the theoretical superstructure of the significance test is a curious restriction. The terminal statement which the test procedure ostensibly endorses provides an answer (if any) devoid of operational value in the context of an experiment rightly undertaken to confirm a positive assertion suggested by prior information. Since the test procedure merely endorses the negation of a null hypothesis conceived within the strait-jacket of the single infinite hypothetical population, the outcome will thus be an irrelevant decision or no decision at all.

The clinical trial sheds light on the legitimate claims of a significance test for another reason. It encourages us to probe

more deeply into the structure of the alleged single infinite hypothetical population from which we extract two groups subjected to equally efficacious treatments. The truth is that the attack rate of a group of persons does not merely depend on what we associate with the treatment criterion. It will depend on age, medical history, nutrition and innumerable other changing circumstances. Thus our assumed infinite hypothetical population is not a fixture; though we may here concede that we can extract a very large number of samples from a factually finite population during a period in which relevant change is negligible. If we merely propose to apply the conclusion which the test endorses to situations in which relevant change is indeed negligible, we shall indeed do no violence to the canon of the fixed historical framework of repetition. Unless new circumstances conspire to make operative an otherwise latent difference with respect to the efficacy of two treatments, we may therefore reasonably continue to believe that their efficacy is constant if previous experience of an infinite hypothetical population conceived in the foregoing terms has justifiably convinced us that this is true; but then we must ask ourselves what we mean by *justifiably* in this context. For the judgment it implies, we assume some rational basis outside the framework of the test procedure dealt with; and the test procedure itself disclaims the title to justify such a conclusion, if the prescribed alternatives are reservation of judgment and rejection of the null hypothesis that our two samples, in Fisher's own words, come from the same infinite hypothetical population.

A new issue arises if we take courage from Fisher's second thoughts (p. 498) and choose a null hypothesis appropriate to the end in view. We then formulate it in terms of a target value  $k$  on the assumption that: (a) we do not wish to relinquish lightly the benefit of substituting treatment B for treatment A; (b) a difference as great as  $k$  is our criterion of minimal efficacy when we use the word *lightly* in this context. With appropriate choice of a one-sided criterion of rejection we may adopt as our null hypothesis either that  $M_d \geq k$  or that  $M_d < k$ . If we choose and reject the latter our equivalent assertion is the former. In the set-up of the Yule-Greenwood

trial, we are then saying that the adoption of treatment B (*inoculation*) will ensure a reduction of the attack rate by at least 100  $k$  per cent, the alternative (treatment A) being *no* inoculation. This at least would seem to be a useful statement, if true; but closer examination raises doubt both about its usefulness and about its credibility.

For the sake of argument we have conceded the claim that the biologist, if convinced that treatment B has no effect in one milieu, may have good grounds for dismissing the possibility that it will be efficacious in a new one; but his legitimate assurance that  $M_d \geq k$  is not on the same footing. Experience has taught him that epidemic diseases may disappear dramatically in response to changes of the social environment without intervention of the sort here signified as treatment. Hence he may well find that the advantage of treatment B has become negligible after lapse of a comparatively short time interval. Though we may define a framework of repetition in terms which plausibly accommodate the formal credentials of the test procedure with the theory of a stochastic model, we are therefore on shifting sands, when we seek to define our framework of repetition in terms of future conduct.

Nor does the mere fact that treatment A in a prophylactic trial of the type dealt with by Yule and Greenwood commonly means *no treatment at all*, restrict the relevance of the difficulty here disclosed. In therapeutic trials now zealously conducted on the same prescription, the dilemma is at least equally real. We have much evidence that: (a) otherwise indistinguishable bacteria may be more or less resistant to the sulphonamides; (b) widespread use of sulphonamides—especially in low dosage—results in selection of resistant strains. Thus experience of sulphonamide therapy which was much more efficacious than  $\text{KMnO}_4$  or  $\text{HgOCN}$  for treatment of gonorrhoea at the time of its introduction in the mid-thirties proved to be highly disappointing when used for British troops (1944) in Italy, where the German authorities had already distributed sulfa-drugs freely as a preventive measure to the prostitute population.

If we dismiss all the foregoing considerations, we have still to dispose of a formidable objection to the use of a statistical

test procedure in the conduct of the clinical trial, and it is an objection which confronts the decision test of Neyman, E. S. Pearson and Wald (Chapter 15) no less than the Yule-Fisher significance test. It is now customary to assume that the problem of the *prophylactic* trial (e.g. whether vaccination is efficacious against attack), is formally identical with that of the *therapeutic* trial (e.g. is penicillin more efficacious than sulphonamides for the treatment of gonorrhoea?) The identification is admissible, only if we lose sight of the end in view. In the practice of preventive medicine our concern is with numbers, and the framework of our problem is essentially one of social accountancy. In the practice of curative medicine our concern is with the sick individual, and unreflective reliance on averages as a criterion of preference may lead us to recommendations inconsistent with the end in view.

That this is so will be immediately apparent if we look at the problem through the spectacles of Claude Bernard (p. 227). Let us therefore consider a fictitious situation. We may suppose: (i) that a disease D is incurable if untreated; (ii) that a clinical trial of the usual type leads us to assess the recovery-rate under treatment A as about 25 per cent and the recovery-rate under treatment B as about 50 per cent. In such circumstances we too easily then content ourselves with a recommendation to step up the recovery rate 25 per cent by substituting treatment B for treatment A. If so, our preoccupation with averages has blinded us to biological realities. If we are alert to the manifold interaction of nature and nurture, the outcome invites us to ask the question: what peculiarities are common to individuals who respectively respond or fail to respond to one or other treatment?

For heuristic reasons, let us now assume that: (a) persons with grey or brown eyes invariably respond to treatment B and fail to respond to treatment A; (b) persons with blue eyes invariably respond to treatment A but do not respond to treatment B; (c) blue-eyed individuals and individuals with grey or brown eyes occur in the population in the ratio 25:50; (d) our clinical trial groups are representative in the sense that the ratio of blue-eyed persons to persons with grey or dark eyes is also close to 25:50. On these assumptions the recovery

rate would be 100 per cent if all D patients with blue eyes continued to receive treatment A and all D patients with grey or brown eyes henceforth received treatment B.

Doubtless, such reflections will not greatly trouble the mind of the true believer. If they do, the true believer may gain what reassurance the circumstances solicit from one of Fisher's most recent pronouncements. He states his matured views on the composition, location and duration of the infinite hypothetical population in terms which are worthy of citation\* because the choice of phraseology suggests that the concept is less the reason for the faith that is in them than the *result* of the application of his method by the militant church of his following:

Briefly, the hypothetical population is a *conceptual resultant* of the conditions we study.

If such considerations are not wholly negligible in the set-up of the therapeutic trial, they are of compelling relevance to the task of the field worker in the social sciences; and it is agreeable to be able to cite one exponent of social statistics alert to the pitfalls which beset reliance on the concept of an infinite hypothetical population. Margaret Jarman Hagood's discussion of its status in *Statistics for Sociologists* (Second Edition, 1947, pp. 429-31) is worthy of citation at length:

In any test of significance there is a testing of some hypothesis about a universe from which the set of observations (a limited universe itself in this case) may be considered a random sample. That is, the logical structure of a superuniverse and the variation expected in random samples from it is the same for the observer sociologist as for the experimentalist. Imagining any experimental counterpart of the logical model is a more difficult matter, however. It is easy enough for the experimentalist to imagine repeated experiments under identical conditions, whether or not he can actually perfect his technique to the degree that he can reproduce conditions identically. His universe of possibilities can therefore be put into meaningful terms; it can at least be imagined, even if it cannot actually be produced. It is not so easy for the sociologist

\* *Proc. Camb. Phil. Soc.*, 22, p. 700.



to imagine a set of observations repeated under conditions identical with those of one date. The fact of change in social and cultural phenomena renders unrealistic any conception of identical repetition of the complex of factors conditioning characteristics such as fertility and level of living. . . . To what, then, does the variation expected from random sampling from such a universe of possibilities correspond? Only a feat of imagination involving an infinite prolongation of a present moment, where conditioning factors remain the same but "chance" factors continue to produce random variation can supply the answer. With this done, the observer sociologist along with the experimentalist still faces the problem of interpretation of the chance variation—with the alternatives of ascribing it to the present limitations in knowledge or to the statistical nature of the occurrence of events. . . . It has been suggested that the limited universe of measures on all of a series of demographic units as of a certain date be considered a sample in time; that the random variations of sampling from a super-universe have their counterpart in the fluctuations which would be observed if we made observations on successive days, or for successive years, while the general influencing conditions would not have altered appreciably. It is evident, however, that such successive fluctuations would not be independent, nor could they be thought of as being produced by forces independent of each other, and therefore they would not be expected to have the same distribution as those produced by chance factors in random sampling (as in fact can be shown to be the case). . . . Another suggestion is that the measures on demographic units may be conceived of as one of an infinite set of such measures secured by dividing the total area surveyed into different series of areal units by shifting of boundaries under certain conditions of contiguity and uniformity of size. The matter of the arbitrary nature of the "lumps" in which our demographic information is secured, and the possible variations to be expected by recombining the information into different lumps has not been explored adequately. While the matter needs attention, it is probably not the answer to the search for a realistic counterpart of the universe of possibilities and to the random variation expected in samples from such a universe. . . . At present, the sociologist must face the fact that the postulated, hypothetical, infinite universe of possibilities, concerning which he tests hypotheses to establish the "significance" of his results, is merely a logical structure, for which he can offer no real counterpart in his research situation. Then what is the utility of such a

construct and of the tests of significance based upon it? The answer to this question is not perfectly clear at the present stage of the application of statistical methods to sociological research.

I shall not comment on Margaret Hagood's concluding remarks concerning the possibility that "a case for the use of such a construct may be made," because it is not the obligation of the research worker to bow to the dictates of statistical theory until he or she has conclusively established its relevance to the technique of enquiry. On the contrary, the onus lies on the exponent of statistical theory to furnish irresistible reasons for adopting procedures which have still to prove their worth against a background of three centuries of progress in scientific discovery accomplished without their aid.

## CHAPTER FIFTEEN

### DECISION, INDECISION AND SAMPLE ECONOMY

IN CHAPTER FIVE (pp. 150-56) we have explored a type of test procedure which embraces model situations to which Bayes's theorem is factually relevant, model situations to which it may be factually relevant for anything we know to the contrary, and model situations to which it can have no factual relevance. Since the theorem of Bayes still casts a pall of irrelevant gloom over controversy concerning the credentials of test procedure, I shall now examine two different current views against the background of a situation which we may approach in accordance with the same three assumptions. Our model will be a fruitfly culture. We shall assume that it contains females of not more than two sorts when classified with respect to all particulars relevant to the question we shall ask, i.e. the alternative hypotheses we propose to test:

- (a) they are normal in the sense that they carry no sex-linked lethal;
- (b) they are heterozygous w.r.t. such a gene.

In accordance with the strictly stochastic modern theory of the gene, we denote the probability that an offspring of a female of type (a) is male by  $p_a = \frac{1}{2} = (1 - q_a)$ , and the probability that an offspring of a female of type (b) is male by  $p_b = \frac{1}{3} = (1 - q_b)$ . We shall assume that we are able to record how many of the  $r$  progeny of females chosen random-wise are males, and we shall seek to formulate a rule of procedure for deciding whether female flies with  $r$  offspring are of type (a) or of type (b) on the understanding that we can set a long-run upper limit (our *uncertainty safeguard*) to the frequency of erroneous assertion, if we follow the rule consistently. If we specify Hypothesis A as the assertion that the female is of type (a) and Hypothesis B as the assertion that the

female is of type (b), our assertion on any single occasion may be:

*either* Hypothesis A true and B false

*or* Hypothesis B true and A false.

Such a rule is a *decision test* in the most literal sense of the term; but it will help us to clarify the credentials of the *significance* test procedure of the opposing school, if we here speak of the former as a *comprehensive* test in the sense that it commits us to an equally definite assertion about the outcome of *every* trial in the assumed framework of repetition. As we have already seen (p. 332), we can make a more restricted type of rule to which we can meaningfully assign an upper limit of statistical uncertainty. For instance, we may say: (a) if the outcome of the unit trial conforms to a prescribed specification, assert that Hypothesis A is false; (b) if the outcome of the unit trial does not conform to the prescribed condition, say nothing at all. We may speak of a rule conceived in these terms as a *partial* test, because the assigned uncertainty safeguard is referable only to assertions about the outcome of a limited class of situations consistent with the sampling process. Since exponents of the significance test have lately (p. 333) found it convenient to accept this limitation, it will be instructive to examine the implications of such a partial test in the context of the model situation specified above before proceeding to formulate a rule which is comprehensive in the sense already defined.

All we have said about our model situation so far is consistent with three possibilities which we shall separately examine:

(a) we know that the culture contains flies of both sorts, and we also know how many of each it contains;

(b) we know that the culture contains flies of only one sort, but we do not know which sort;

(c) for anything we know to the contrary, the culture may contain flies of both sorts or of one sort only.

To define any decision rule—comprehensive or partial alike—in terms consistent with the argument of Chapters Four

and Five, we must specify in advance both an acceptable uncertainty safeguard and the size of the class of samples subsumed by the test procedure as a basis for the definition of our criterion of decision—acceptance and rejection, if comprehensive, or rejection and reservation of judgment, if partial. For illustrative purposes we shall here restrict our discussion to individual female fruitflies with 22 offspring. The probability that the number of males in an  $r$ -fold fraternity will be  $x$  is:

$$r_{(x)}P_a^x(1 - p_a)^{r-x} = 22_{(x)}2^{-22}, \quad \text{if Hypothesis A} \\ (p_a = \frac{1}{2}) \text{ is true}$$

$$r_{(x)}P_b^x(1 - p_b)^{r-x} = 22_{(x)}2^{22-x} \cdot 3^{-22}, \quad \text{if Hypothesis B} \\ (p_b = \frac{1}{3}) \text{ is true.}$$

From tables of the Binomial Distribution prepared by Dr. Churchill Eisenhart and his co-workers\* we thus obtain the following data relevant to the types of test prescription we shall now explore w.r.t. 22-fold samples, denoting respectively by  $P_a$  and  $P_b$  the probabilities assigned by the hypotheses A and B to a specified range of  $x$ :

$$\sum_{16}^{22} 22_{(x)} \cdot 2^{-22} = P_a(x > 15) \quad \sum_{16}^{22} 22_{(x)} \cdot 2^{22-x} \cdot 3^{-22} = P_b(x > 15) \\ = 0.0262 \quad \simeq 0.0002$$

. . . . .

$$\sum_7^{15} 22_{(x)} \cdot 2^{-22} = P_a(6 < x \leq 15) \quad \sum_7^{15} 22_{(x)} \cdot 2^{22-x} \cdot 3^{-22} = \\ P_b(6 < x \leq 15) \\ = 0.9476 \quad \simeq 0.6380$$

. . . . .

$$\sum_0^6 22_{(x)} \cdot 2^{-22} = P_a(x \leq 6) \quad \sum_0^6 22_{(x)} \cdot 2^{22-x} \cdot 3^{-22} = P_b(x \leq 6) \\ = 0.0262 \quad = 0.3618$$

We may thus arbitrarily split our range of sample score

\* *Tables of the Binomial Probability Distribution*. U.S. Dept. of Commerce, National Bureau of Standards, Applied Mathematics Series No. 6. Washington, 1949.

# STATISTICAL THEORY

values (0, 1 . . . . 21, 22 males) with probabilities assignable on the alternative prescribed hypotheses as below :

## On Hypothesis A

0.0262	0.9476	0.0262
0	6.5	15.5
		22

## On Hypothesis B

0.3618	0.6386	$\simeq 0.0002$
0	6.5	15.5
		22

0.3618	0.6382
0	6.5
	22

Our sole reason for making the split in this way is that we can examine the implications of the Fisher test theory without relinquishing the 5 per cent feeling it confers. On Hypothesis A, the mean of the distribution is 11 and the range of scores from 7 to 15, i.e. all admissible scores in the range  $11 \pm 4.5$ , defines a class of samples which conform to the 2-fold condition that they: (a) lie symmetrically about the mean; (b) include nearly 95 per cent of those which we meet in an endless sequence of randomwise trials. In the situation before us there is no compelling reason why we should prefer Hypothesis A to Hypothesis B as our null hypothesis in Fisher's sense, since neither is ambiguous and one is just as tractable as the other in terms of algebra or computation. If we here choose Hypothesis A, we do so because: (i) most laboratory workers brought up on the good books are apt to identify the null hypothesis with the more conservative hypothesis; (ii) neither Fisher nor his disciples clearly specify the criterion for preferring a modular (two-sided) criterion of rejection

rather than a vector (one-sided) when the distribution happens, as is prescribed by Hypothesis B, to be skew.

If we now proceed as we should proceed in accordance with the prescribed significance test drill, we have already made an important concession to the Forward Look by stating our rule in the form:

- (a) reject Hypothesis A if the score lies outside the range 7 to 15 inclusive *for the 22-fold sample*,
- (b) otherwise reserve judgment.

The rule so stated prescribes that we shall reject the null hypothesis when true only in 5.2 per cent of the situations we encounter; and this may sound like the same thing as saying that we reject it at the 5.2 per cent significance level; but by stating the rule as above, we have relinquished the right to look at the sample before we specify the score rejection criterion. Contrariwise, the customary significance test procedure endorses our right to say that we shall reject the null hypothesis, if:

- (a) the score is  $X_r = \pm (x - M)$  if the sample contains  $x$  males

- (b)  $P(M \pm \overline{x - 1}) \geq 0.948$ , i.e. a deviation from the mean of the distribution numerically *as great as*  $X_r$  has a probability of 0.052 for samples of the same size as the observed one.

In accordance with decision test procedure, we here say that *not more* than 5.2 per cent of our assertions will be false *about samples of 22 offspring* if we reject the null hypothesis (here Hypothesis A) when  $x$  is less than 7 or more than 15. Thus we have implicitly shouldered the obligation to specify the size of the sample, and hence a score rejection criterion *before* we examine the evidence. At this stage, some reader may regard this as a distinction without a difference; but a fuller examination of our model situation will disclose an eminently practical reason for insistence on preliminary specification of sample size other than an unduly fastidious concern for the statement of a formula consistent with the Forward Look.

First let us examine our rule against the background of the alternative admissible assumptions on the understanding that our rule permits us only to say that Hypothesis A is false and otherwise to reserve judgment. If Hypothesis A is actually true, this means that roughly 5 per cent of our assertions will be false; but if Hypothesis B is true, none of our statements will be false, since we have excluded our right to say that Hypothesis A is the correct one. Thus at most, not more than about 5 per cent (more closely 5.2 per cent) of any statements we make within the framework of the rule can be erroneous.

Before scrutinising more closely a so seemingly gratifying conclusion, we shall do well to notice *en passant* that we have picked up an unnecessary encumbrance in our wanderings within the shadowy domain of the infinite hypothetical population. We are much less likely to meet scores of over 15 if Hypothesis B is true than otherwise. All that is relevant to the rule of decision to which we seek to attach an uncertainty safeguard at an acceptable level (here assumed to be 5.2 per cent) is that the number of males in the sample will be greater in the long run if Hypothesis A is true than it will be if Hypothesis B is true. Why therefore should we carry over from the Gaussian domain the inclination to use a 2-sided rejection criterion referable to an irrelevant mean value? Why should we not more boldly state that we shall reject Hypothesis A if  $x < 6$  and otherwise reserve judgment? Since  $P_f \simeq 0.026$ , we are then entitled to attach an upper limit of approximately 2.6 per cent to the uncertainty safeguard of the rule.

Let us now look at the issue, as we have seen from the citations on pp. 334-335 that the consumer customarily does look at it, and with every encouragement from the good books. We may reasonably assume that we need not go to great trouble to gratify the whims of a particular school of statistical theory without the assurance that it provides a recipe for arriving at a decision one way or the other. We shall accordingly modify the statement of the foregoing by substituting the words *accept Hypothesis B* for the words *reserve judgment*. We must then adopt a one-sided rejection criterion, and may formulate a comprehensive rule in the following terms:



- (i) reject Hypothesis A (and hence accept Hypothesis B)  
if  $x \leq 6$ ,
- (ii) accept Hypothesis A (and hence reject Hypothesis B)  
if  $x > 6$ .

In the terminology of Neyman and of E. S. Pearson (*Proc. Camb. Phil. Soc.* 29, 1933) we may here make errors of two kinds:

A. We may reject Hypothesis A when it is true, thus accepting Hypothesis B when it is false with conditional probability  $\alpha$ .

B. We may accept Hypothesis A when it is false, thus rejecting Hypothesis B when it is true with conditional probability  $\beta$ .

When we designate Hypothesis A as the null hypothesis, the sample score as  $x$  and the rejection score criterion as  $x_r$ , we may write:

$P_{x.a}(x \leq x_r) = \alpha$  as the conditional probability of making  
an error of the first kind

$P_{x.b}(x > x_r) = \beta$  as the conditional probability of making  
an error of the second kind.

For the comprehensive rule last stated,  $x$  is the number of males,  $r = 22$  and we set:

$$x_r = 6; P_{x.a}(x \leq 6) = \alpha = 0.0262; P_{x.b}(x > 6) = \beta = 0.6382$$

We may cover all three possibilities embraced by picking out from our culture *at random* female fruitflies with 22 offspring, if we postulate that: (a) the proportion of normal females is  $P_a$ , that of lethal carriers being  $P_b = (1 - P_a)$ ; (b) the value of  $P_a$  lies in the range 0 — 1 inclusive, being unity when all the flies are normal and zero if all are carriers. The probability of all possible results of the test procedure then follows from the multiplication theorem, viz.:

- (i) Hypothesis A is true and we accept it:  $P_a(1 - \alpha)$
- (ii) *ditto* but we reject it:  $P_a \cdot \alpha$
- (iii) Hypothesis B is true and we accept it:  $P_b(1 - \beta)$
- (iv) *ditto* but we reject it:  $P_b \cdot \beta$

By the addition theorem we thus derive:

Probability of *True* assertion:

$$P_t = P_a(1 - \alpha) + (1 - P_a)(1 - \beta) = 1 - \beta + P_a(\beta - \alpha) \quad (i)$$

Probability of *False* assertion:

$$P_f = P_a \cdot \alpha + (1 - P_a)\beta = \beta + P_a(\alpha - \beta) \quad . \quad (ii)$$

In the foregoing situation  $\alpha = 0.0262$  and  $\beta = 0.6382$ ,  $(1 - \beta) = 0.3618$  and  $(\alpha - \beta) = -0.6120$ , so that

$$P_f = 0.6382 - P_a(0.6120)$$

We may now tabulate the uncertainty safeguard of the rule for different possible proportions of normal flies in the culture, i.e. different prior probabilities which we can assign to the null hypothesis in accordance with our assumption that the culture contains each of these two sorts of female fruitflies, one chosen randomwise for each test performance

$P_a =$	$P_f \simeq$
0.999	0.027
0.990	0.032
0.900	0.087
0.500	0.332
0.200	0.516
0.050	0.608
0.001	0.638

We see from this what is formally evident from (ii):

(i) As  $P_a$  approaches unity, i.e. when nearly all the females conform to the specification of the null hypothesis (Hypothesis A),  $P_f$  approaches  $\alpha$  more and more closely;

(ii) as  $P_a$  approaches zero, i.e. when nearly all the females are lethal carriers and thus do *not* conform to the null hypothesis,  $P_f$  approaches more and more closely to  $\beta$ .

It is now easy to state the result of the test procedure, when we conceive the situation in terms of *sampling in one stage*: i.e. we know that the culture contains only one sort of fly, but

not which sort. This is the situation in which we commonly find ourselves, when we talk about testing the truth of a hypothesis. If Hypothesis A is true, we cannot make an error of the second kind and  $P_j = \alpha$ . If Hypothesis B is true, we cannot make an error of the first kind and  $P_j = \beta$ . All we can say is that  $\beta \geq P_j \geq \alpha$  (if  $\beta > \alpha$ ) or  $\alpha \geq P_j \geq \beta$  (if  $\alpha > \beta$ ). This interpretation of the rule covers the case which arises when we do not know whether the culture contains both sorts of fruitflies or one only, since we then assume that  $P_a$  can have any value in the range 0 to 1 as above.

The reader will now see why it is necessary for the exponents of the significance test conceived as such in terms of a single null hypothesis without reference to any alternative must needs insist that the outcome of their test prescription does not entitle the performer to *accept* it. In this case the adoption of a rejection criterion, which confers on a *partial* decision rule with this restriction an uncertainty safeguard  $\alpha \leq 0.026$ , entitles us to say no more about the result of operating the corresponding comprehensive rule, i.e. acceptance of the null hypothesis when we do not reject it, than that nearly 64 per cent of the assertions the rule endorses might be false in the long run.

To operate a comprehensive decision rule which takes stock of admissible alternative hypotheses, we have therefore to choose our rejection criterion with due regard to the conditional probability ( $\beta$ ) of error of the second kind as well as to the so-called significance level ( $\alpha$ ) referable to the error of the first kind. From inspection of (ii) we see that  $P_j = \alpha$ , if  $\beta = \alpha$ , and we cannot decrease  $\beta$  for a prescribed sample size unless we increase  $\alpha$ . For example, our tables show that  $P_{x.a}(x \leq 9) \simeq 0.462$  and  $P_{x.b}(x > 9) \simeq 0.162$  for 22-fold samples. If we choose to reject the null hypothesis when the score (male offspring) is less than 10, we shall reject Hypothesis B if true in only about 16 per cent of samples which we shall then encounter; but we shall now reject Hypothesis A if true in about 46 per cent of samples we shall then encounter. To prescribe a comprehensive test procedure consistent with a preassigned acceptable uncertainty safeguard we must thus define our sample size ( $r$ ) in advance.

This will be a laborious procedure, if we use the tables of

the binomial, but we can define a standard score of unit variance with an approximately normal distribution if  $r$  is fairly large (see p. 160), i.e. for the hypotheses under consideration  $r > 30$ . If  $p_h = (1 - q_h)$  is the long-run proportion of male offspring on hypothesis  $H$ , and  $x$  is the actual number in the  $r$ -fold sample, the corresponding square standard score is

$$\frac{(x - r \cdot p_h)^2}{r \cdot p_h \cdot q_h} = c_h^2$$

In the situation we here examine for illustrative purposes, we have

$$p_a = \frac{1}{2} ; c_a = \frac{2x - r}{\sqrt{r}} \quad . \quad . \quad . \quad (iii)$$

$$p_b = \frac{1}{3} ; c_b = \frac{3x - r}{\sqrt{2r}} \quad . \quad . \quad . \quad (iv)$$

If we wish to make  $P_f = \alpha$ , in which event  $\alpha = \beta$ , we shall choose our rejection score criterion  $(x_r)$  so that  $c_b = -c_a$ , whence

$$x_r = \frac{(1 + \sqrt{2})r}{3 + 2\sqrt{2}} \simeq 0.414r \quad . \quad . \quad . \quad (v)$$

Let us then assign as our acceptable uncertainty safeguard  $P_f \leq 0.05$ . The table of the integral shows that 0.05 of the normal curve lies in the region from  $-\infty$  to  $c_h = -1.64$  or from  $c_h = +1.64$  to  $+\infty$ . Whence by substitution in (iii), we get

$$\sqrt{r} = \frac{1.64}{0.182} \simeq 9.01$$

$$r \simeq 81.2$$

Also from (v),  $x = 33.9$ . We shall thus guarantee  $P_f < 0.05$  if we confine our attention to samples of 82 and:

reject Hypothesis A in favour of Hypothesis B if  $x < 34$

reject Hypothesis B in favour of Hypothesis A if  $x > 33$

By examining only samples of larger size, we can of course

reduce the upper limit of uncertainty. As before, we have only to choose our score rejection criterion to satisfy the foregoing criterion  $\alpha = \beta$ , so that  $-c_a = c_b$ .

At this stage, the reader may ask the following question: is there any essential difference between a partial decision rule and a significance test in the Yule-Fisher tradition? One answer to this is that a decision test of any sort, as we have defined it, imposes on us the obligation to state in advance the size of the sample to which the test prescription applies. I have elsewhere contended that such a consistently behaviourist interpretation of the test procedure as a rule of conduct does indeed impose on us this obligation. The laboratory worker, not as yet restored to self-confidence after the traumatic discovery that he can no longer accept with propriety a null hypothesis seductively irrelevant to his main preoccupations, and therefore all too ready for belated assurance concerning the utility of the test prescription, may be less impressed by such formal considerations than by the practical consequences of performing tests without a preliminary examination of the bearing of sample size on any realisable reward for much expended arithmetical toil.

If anxious to find a formula for compromise, we may indeed gracefully yield to the temptation to state that we make an error of the second kind only if we accept the null hypothesis when false, and hence that we cannot make the error of the second kind within the framework of a rule to reserve judgment when our sample score does not lie outside the chosen limit or limits definitive of  $\alpha$ , i.e. the conditional probability of wrongly rejecting the null hypothesis. We may therefore all too readily welcome the disclosure that we can at least say something directly relevant to the truth or falsehood of our decisions, if we restrict the verdict to rejection and reservation of judgment. At its face value, this contention is admissible; but it entails devastating consequences to highly publicised claims which have been largely responsible for the popularity of the significance test formula.

It has been the plea of R. A. Fisher and of his followers that the tests elaborated by them have special advantages on grounds of economy; and an imposing edifice of *small sample*

*theory* has arisen on this understanding. Indeed, some devotees have lapsed into the idiom of poetic diction appropriate to the magnitude of their benefits so conceived. In one context, we hear that they are "exquisitely sensitive." In another, Darlington is content to use the idiom of ballistics with pardonably patriotic exaltation, when he assures us that:

Modern statistical methods, largely developed by Englishmen, have transformed our knowledge of how to extract information from numbers. They have become in recent years one of our most powerful and most general instruments of discovery. Our great Government departments are busy collecting for us numbers, so-called statistics, on a vast scale. . . . Modern statistical methods would demand new activities of explosive violence.

What the school of R. A. Fisher defines as the appropriate criterion of *economy* in the context of small sample theory need not detain us, because the concepts of efficiency and sufficiency\* as Fisher first defined them, emerged in the domain of *point*-estimation. In any event, we can presume to know what the consumer will reasonably expect of it. If told that a test is economical in its demand on sample size, he or she will reasonably expect that its design will commonly guarantee the possibility of arriving at the only sort of decision it does endorse; but our foregoing model situation shows that this is a hope forlorn. If all our fruitfly mothers of 22 offspring are lethal carriers, the decision to reserve judgment or to reject the null hypothesis at the 2.6 per cent significance level on the basis of a one-sided criterion, and approximately 5 per cent if we adopt a 2-sided criterion, signified that 64 per cent of all the tests we shall carry out will be indecisive.

It is revealing to consider how the Fisher test procedure works out, when the size of the fruitfly fraternity is 9. We need not here rely on the tables. In the foregoing symbolism

$$(i) P_{x.a}(x \leq 1) = \frac{1 + 9}{2^9} < 0.02$$

$$(ii) P_{x.b}(x > 1) = 1 - \frac{11 \cdot 2^8}{3^9} > 0.85$$

\* R. A. Fisher (1922). "*Foundations of Theoretical Statistics.*" *Phil. Trans. Roy. Soc. A.*, ccxxii.

In words, what this means is that: (i) the probability of meeting 9-fold samples containing 0 or 1 males is less than 0.02 when Hypothesis A is true; (ii) more than 85 per cent of all samples we shall meet will actually contain as many as 2 males if Hypothesis B is true. On this understanding, we may therefore frame a partial decision rule referable to a unique null hypothesis in the following terms: if the 9-fold sample contains less than two males, reject Hypothesis A and otherwise reserve judgment if it contains 2 or more males. We may then admittedly assign an uncertainty safeguard  $P_j < 0.02$  to any positive assertions to which we commit ourselves. The fact remains that Hypothesis B may be true, in which event more than 85 per cent of our samples will contain 2 or more males. If, therefore, Hypothesis B is true, our rule enjoins on us to reserve judgment about 85 per cent of the samples we meet. To sidestep the error of the second kind by a verbal afterthought, we thus arrive at the alarming conclusion that the test prescription is consistent with the possibility that the result of applying it will be inconclusive in about 85 per cent of all opportunities we may have for doing so.

Thus rejection of the null hypothesis at the one-sided 2 per cent (or at the two-sided 4 per cent) significance level means that we may have nothing positive to show for the labour of performing some 85 per cent of all tests we carry out. In this context, the labour is admittedly a trivial issue; but it is by no means a trivial undertaking to ring the changes on the test battery of an Analysis of Covariance. The investigator who shoulders the obligation to do so, and accordingly relinquishes his own prerogative of experimental design to fit the assembly of data in the mould prescribed by the test prescription, must therefore reluctantly abandon the hope that a significance test accomplishes what little R. A. Fisher (p. 500) claims for it in his later pronouncements *ex cathedra*.

In any event, it is now clear that the consumer will not know how much of the effort is fruitless unless he can give a more precise meaning to Snedecor's statement (p. 335) expressing a somewhat lax view of the implications of rejection, viz. that the relevant parameter is "so small as to be unimportant to his investigator." We can indeed do this only if

# STATISTICAL THEORY

we set an upper limit to what we deem to be small. If our null hypothesis is  $p_a = \frac{1}{2}$ , we shall then say that we ask the test to guarantee that we shall rarely suspend judgment when

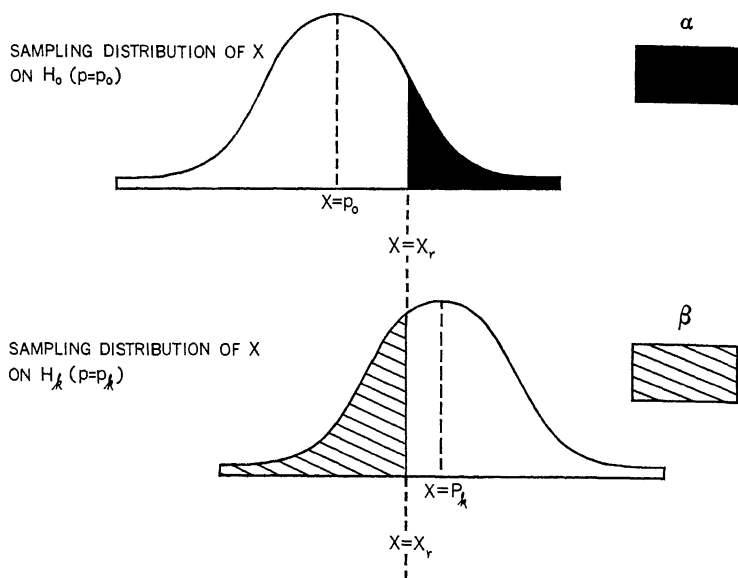


FIG. 1.—Decision—partial and comprehensive

*Above.*—Sampling distribution of a score  $x$  in accordance with the Hypothesis  $H_0$  that  $p = p_0$ .

*Below.*—Sampling distribution of a score  $x$  in accordance with the Hypothesis  $H_k$  that  $p = p_k$ .

*Rejection Criterion*

If  $x > x_r$

If  $x \leq x_r$

$\alpha$

$\beta$

*Partial Rule*

Say  $H_0$  is false

Say nothing

Probability of rejecting  $H_0$  when  $H_0$  is true

Probability of suspending judgment when  $H_k$  is true

*Comprehensive Rule*

Say  $H_0$  is false and  $H_k$  is true  
Say  $H_k$  is false and  $H_0$  is true  
Probability of rejecting  $H_0$  when  $H_0$  is true  
Probability of rejecting  $H_k$  when  $H_k$  is true

the true value ( $p_b$ ) of the parameter has some limiting value ( $k$ ) which we may define in a one-sided or two-sided way as:

- (i)  $p_b < k_2$ ; (ii)  $k_2 > p_b > k_1$ ; (iii)  $p_b > k_1$



Thus we cannot confer any intelligible connotation on the economy of the test procedure from the viewpoint of the consumer who has nothing to show for a non-committal test result unless we can set up some clear-cut alternative to the hypothesis we propose to discredit. In the fruitfly model situation discussed above, only one alternative to the null hypothesis is admissible; but the issue raised by Snedecor demands a different formulation in the set-up of the prophylactic trial. If we denote by  $p_a$  and  $p_b$  respectively attack (or mortality) rates for treatments A and B, having reason to expect that treatment B is the more efficacious, our concern will be to decide whether the difference ( $p_a - p_b = M_d$ ) attains a certain target value ( $k$ ), i.e.  $M_d \geq k$ . To be sure, this hypothesis is ambiguous in the sense that it is consistent with an infinitude of sampling distributions, but we can dispose of this objection if we proceed as follows.

We shall suppose that our null hypothesis ( $H_k$ ) is  $M_d = k$ , and that we shall reject it when the observed attack rate difference ( $d$ ) is less than  $d_r$ , the latter being itself less than  $k$ . Our error of the first kind will be

$$P_{d,k}(d < d_r) = \alpha$$

For any other hypothesis ( $H_m$ ) such that  $M_d = m > k$ , we may write:

$$P_{d,m}(d < d_r) < \alpha$$

If  $\alpha$  is the probability that we shall reject the hypothesis  $M_d = k$  when it is true, the probability that we shall reject the range of hypotheses  $M_d \geq k$  is thus  $P_f \leq \alpha$ . Hence the hypothesis  $M_d \geq k$  is unambiguous if we: (a) choose a *vector* rejection score criterion to satisfy the acceptable level of uncertainty  $P_f = \alpha$  when  $M_d = k$ ; (b) content ourselves with stating the uncertainty safeguard of the rule in the form  $P_f \leq k$ . On this understanding, we can view the claims of small sample theory from a new aspect, if we take a *back stage* view of a  $2 \times 2$  table (proportionate score difference or so-called 1 d.f. Chi Square test) prescribed by Fisher for trials of the sort discussed by Yule and Greenwood and indeed used by the latter.

Accordingly, we shall assume: (i) that we know the true value of  $p_a$  which is the attack rate in an infinite population specified by treatment A; (ii) that we agree about the target value  $k$  which defines  $p_b = (p_a - k)$  as that of a population specified by treatment B; (iii) that  $k$  so specified is the least operational advantage we choose to regard as important. In the last expression  $(p_b - p_a) = 0 = k$  if the two populations are identical in accordance with the Yule-Fisher test prescription.

We may then define the variance of the distribution of an observed proportionate difference  $d = (p_{a.s} - p_{b.s})$  referable to observed attack rates  $p_{a.s}$  and  $p_{b.s}$  of equal ( $r$ -fold) samples as:

$$\begin{aligned}\sigma_d^2 &= \frac{p_a(1 - p_a)}{r} + \frac{p_b(1 - p_b)}{r} \\ &= \frac{2p_a(1 - p_a) - k(1 - 2p_a + k)}{r} \quad . \quad (vi)\end{aligned}$$

If both  $p_a$  and  $p_b$  lie in the range  $0.4 - 0.6$ , the normal curve of unit variance will describe adequately for our purpose the distribution of the standard score of 50-fold samples:

$$c_k = \frac{(d - k)}{\sigma_d} \quad . \quad . \quad . \quad . \quad . \quad (vii)$$

Since we are here taking a back-stage view by assuming that we have a firm value  $p_a$  for treatment A, we may set  $p_a = 0.5$  for illustrative purposes and shall then define a square normal score of unit variance (i.e. a Chi Square variate of 1 d.f.) by

$$c_k^2 = \frac{2r(d - k)^2}{1 - 2k^2}$$

We shall now consider the alternative hypotheses: (a)  $H_0$  if  $k = 0$ ; (b)  $H_k$ , if  $k = 0.10$ . In words, this means: (a) the true attack rates for both treatments are 50 per cent, the two treatments thus being equally efficacious on  $H_0$ ; (b) the attack rates for treatments A and B are respectively 50 per cent and 40 per cent, treatment B being therefore 10 per cent more

efficacious than treatment A on  $H_k$ . We shall then write the standard scores of the two sampling difference distributions as :

$$c_0 = + d\sqrt{2r} ; c_k = \frac{(10d - 1)\sqrt{r}}{7} . \quad (\text{viii})$$

Since a large positive difference  $d$  will be more rare than otherwise if  $H_0$  is true, we shall choose a one-sided rejection criterion  $d_r > 0$  such that  $P_{d,0}(d > d_r) = \alpha$  and set  $\alpha = 0.05$ . A partial decision rule is then rejection of  $H_0$  if  $d > d_r$  with an uncertainty safeguard  $P_r = 0.05 = \alpha$ . Accordingly, we choose  $d_r$  so that  $d_r = d = (p_{a,s} - p_{b,s})$  when  $c_0 = + 1.64$ , since 5 per cent of all sample values lie in the range of standard scores ( $c_k$ ) from  $+ 1.64$  to  $+\infty$ . We have then fixed  $c_k$  in (viii) for any specified value of  $r$  and are able to give a more definite form to the issue Snedecor states in the passage cited. Our rejection criterion means that we suspend judgment when  $c_0 < + 1.64$ ; and if  $r = 50$  this means that  $d < 0.164$  and  $c_k < 0.65$ . If we suppose that  $H_k$  is true, we shall thus suspend judgment whenever the score deviation  $(d - k)$  lies in the range from  $-\infty$  up to  $+ 0.65$ , as is true of 74 per cent of the samples that we shall encounter. Had we chosen as our target value of minimum acceptable efficacy 5 per cent advantage ( $k = 0.05$ ), we should obtain  $c_k = + 1.14$ . This bounds 87 per cent of the area of the normal distribution. Thus we shall suspend judgment about 87 per cent of samples we encounter, if we carry out the usual drill at the 5 per cent significance level on equal samples of 50 when there is a real advantage of 5 per cent.

By defining  $k$  as our target value of minimal efficacy, we have said that any difference  $M_d = (p_a - p_b)$  less than  $k$  is unimportant, or, in Snedecor's own (p. 335) words "so small as to be unimportant for our investigation." We thus arrive at the following conclusions with reference to the Fisher Chi Square test for the  $2 \times 2$  table of the clinical trial :

- (i) if we interpret the alternative to rejection in the sense endorsed by Snedecor, the overwhelming majority of our judgments may be wrong :

(ii) if we stand fast by Fisher's injunction to suspend judgment, an overwhelming majority of the tests we carry out may be indecisive *vis-à-vis* the only type of decision which justifies recourse to the procedure.

In any event, we shall not know how much expended effort may well be fruitless, unless we explore what Neyman calls the *power* of the test procedure. This we shall now define. Having chosen some unique hypothesis  $H_0$  referable to a unique parameter  $p_0$  as the basis of a partial decision rule or significance test rehabilitated in such terms, we are free to consider what would happen if any hypothesis  $H_k$  referable to the corresponding parameter  $p_k$  is admissible in the range  $p_k > p_0$ . We must then preassign a score rejection criterion  $x_r$ , so that  $P_0(x > x_r) = \alpha$  defines an acceptable uncertainty safeguard against *wrongly* rejecting  $H_0$ . If any admissible alternative hypotheses  $H_k$  is true the probability that we shall suspend judgment is  $P_k(x < x_r) = \beta_k$ , being numerically identical with the error of the second kind, if we operate the corresponding comprehensive test. The power of the test is then  $(1 - \beta_k)$ , being the probability that we shall reject  $H_0$  when  $H_k$  is true.

For illustrative purposes we may return to the foregoing discussion of a normally distributed difference  $(p_a - p_b) = k$ . It will suffice to consider two admissible alternatives to the null hypothesis  $H_0$ , viz.:

$H_0$	that	$k = 0$
$H_1$	that	$k = 0.05$
$H_2$	that	$k = 0.10$

By using (viii) above, we then obtain for  $c_0 = 1.64$  so that  $\alpha = 0.05$  the following values of  $(1 - \beta_k)$

$r$	$H_1$ true	$H_2$ true
50	0.13	0.26
100	0.18	0.45
200	0.26	0.64
400	0.41	0.88

. . . . .

*Ceteris paribus*, the power of the test increases as we increase the size ( $r$ ) of the sample or increase  $k$ . If we postulate that  $k = e$  is positive but infinitesimal, we shall require indefinitely large samples to ensure a high power test procedure in Neyman's sense; but this form of words is liable to conceal the essential lesson the table of the so-called power function conveys. If the value of  $(1 - \beta_k)$  is high, that of  $\beta_k$  is low and *vice versa*. To say that the power of a test is low for any sizeable specification of the sample thus means that  $\beta_k$  is high; but if  $H_k$  is true,  $\beta_k$  is the proportion of indecisive partial decision tests we shall perform on the basis of the null hypothesis  $H_0$ . We thus reach the following conclusion: if the null hypothesis ( $k = 0$ ) is false, no specification of sample size, however large, can guarantee the opportunity of arriving at a definite decision as the outcome of an appreciable proportion of test applications.

One hopes that the force of the foregoing considerations will commend themselves to any reader who has not a heavy emotional investment in the performance of the significance test ritual. If so, some will have begun to wonder: (a) how the plea of small sample economy advanced by Fisher's disciples gained such widespread assent; (b) why there has been so belated a recognition of logical issues within easy grasp of the consumer with little mathematical knowledge. The first question claims attention, if only because intellectual modesty forbids us to dismiss lightly the views of others possibly more clever than ourselves. When we have arrived at contrary convictions by a different route, it is therefore seemly to ask how misunderstanding may well have arisen.

The source of misunderstanding in this setting is not far to seek, and illustrates a disposition we have already recognised as a perennial source of confusion. When the theoretician speaks of the rigorous proof of a theorem, we should be at all times alert to the distinction between: (i) clarification of the logical credentials of a statistical procedure in the domain of action; (ii) refinement of a cognate algebraic technique in conformity with the requirements of the pure mathematician. If we examine the sense in which Fisher gave the first rigorous proof of the  $t$ -distribution, we have an answer to the first of the two questions stated above. Let us therefore do so.

The null hypothesis of the significance test procedure commonly invokes the assumption that the unit sample distribution of the universe of choice is normal. In fact it never is, and never can be; but the assertion may be sufficiently near to the truth to justify recourse to an appropriate sampling distribution referable thereto. If so, we may say that the distribution of the mean of  $r$ -fold samples from a universe whose definitive parameters are  $M$  and  $\sigma$  is normal with definitive parameters  $M$  and  $\sigma_m = \sigma \div \sqrt{r}$ . Now in practice, we do not know the true value of  $\sigma$ ; but we may be content to assume that the sample root-mean-square deviation  $s$  is a sufficiently reliable estimate of  $\sigma$  and that  $s_m = s \div \sqrt{r}$  is accordingly a sufficiently reliable estimate of  $\sigma_m$  if  $r$  is large enough. Practitioners in the domain of the theory of error from Gauss well into the first half of the subsequent century consistently made the best of a bad job in this way when they cited the precision index  $h$  or probable error of a point estimate; and the school of K. Pearson adopted the practice when applying the theory to a new class of problems.

Under the pen-name *Student*, W. H. Gossett (1908) published what is a purely algebraic theorem to the effect that the distribution of the ratio of the deviation of the sample mean from its true value to the estimated standard deviation ( $s_m$ ) is definable for a normal universe without reference to the unknown parameter  $\sigma$  for all values of  $r$ . At the time, this attracted little attention, partly because the author's proof is fallacious, and perhaps partly also because his symbolism is very obscure. In 1923 Burnside (*Proc. Camb. Phil. Soc.*, Vol. 21) took up the issue for the Gaussian domain in a paper which explicitly refers to, and disclaims, what we have elsewhere (p. 193) called the Legendre-Laplace axiom: "it would be difficult to justify the assumption that because a particular value of the precision constant makes the probability of the observed event as great as possible, the precision constant necessarily has that value." The reader will recall that the precision constant  $h$  in this context is  $(\sigma\sqrt{2})^{-1}$  in the symbolism of statistical tests; and on that understanding Burnside derives a distribution closely related to the  $t$ -distribution of Gossett.

In the same year and in the same journal, the publication

of Burnside's paper prompted R. A. Fisher (1923) to direct attention to the neglect of "the brilliant work of Student," asserting that it "is so fundamental from the theoretical standpoint and has so direct a bearing on the practical conclusions to be drawn *from small samples*." Fisher's communication has attracted more attention than that of Burnside for two reasons, and with some justice.\* He announced his intention of publishing, as indeed he did subsequently publish, a table of the integral with a foolproof explanation of how to use it for the benefit of consumers with no mathematical pretensions to understand its derivation. He also invoked a new algebraic technique cognate to contemporary preoccupation with the theory of relativity, and later exploited the same method to derive distributions of other statistics—notably the ratio of different variance estimates—referable to sampling in a putative normal universe.

In the sense that the distributions of the  $t$ -ratio and of other sample statistics of the Fisher test battery do not depend on the unknown  $\sigma$  of the putative normal universe, they are equally *exact* for samples of any size; and we may invoke them without the reservation that they are valid only for large samples, if we have good reason to invoke them at all. When we say that the Fisher test prescriptions are valid for small samples, all we therefore rightly mean is that the theoretical distributions on which they rely are valid in the foregoing sense; but this assertion has no direct bearing on the claim that such tests are intrinsically economical in the sense that they satisfy the consumer's demand for a decisive outcome of their use.

Let us now turn to the second question stated above. How

\* In a review (*Scientific Monthly*, LXII, 1951) of the memorial volume referred to in *Appendix IV*, Neyman delivers what may well be the verdict of the next generation in the following terms:

A very able "manipulative" mathematician, Fisher enjoys a real mastery in evaluating complicated multiple integrals. In addition, he has a remarkable talent in the most difficult field of approaching problems of empirical research. As a result, his lifework includes a series of valuable contributions giving exact distributions of a variety of statistics, such as the correlation coefficient, the central  $\chi^2$  with due allowance for degrees of freedom, the noncentral  $\chi^2$ , the quotient of two  $\chi^2$  etc. etc. These distributions are bound to stay on the books and be used continuously.

can we explain the belated recognition of considerations first advanced by Neyman and E. S. Pearson more than twenty years ago? We may seek an answer to this on more than one level. Here we shall consider two. While it seems clear that the view of test procedure elaborated by Neyman and E. S. Pearson is traceable to the period in which tests recommended by R. A. Fisher were still widely accepted as novel, we must also recall that the only claim of the latter to novelty is the invocation of new algebraic techniques to specify more or less relevant advantages of invoking particular sample distributions. In fact, they signalise no essential break with a long tradition transmitted by Quetelet through Galton to Pearson and incorporated in postulates defined in all essential particulars by Yule and Greenwood. Inescapably, therefore, the emergent concept of a decision rule had to fit as best it could into a pre-existing framework of custom thought.

Thus the concept of test power which usefully calls attention to our obligation to define the size of the sample, if we wish to confer any intelligible rationale on a test procedure, succumbed to a period of arrested development. To woo their elders and contemporaries, Neyman and Pearson almost succeeded in disposing of an unwanted child by a concession which invalidates reasonable grounds for a hopeful attitude towards its survival. They introduced the concept of a unique *uniformly most powerful* (U.M.P.) test, i.e. a partial decision test which is as powerful as possible in the sense defined above. For the reason mentioned above, this was a decisively backward step, since the most powerful test definable cannot necessarily satisfy one presumptive demand of the consumer, i.e. that he will commonly be able to arrive at a definite decision. Till Wald, as a newcomer in the field, interpreted the power concept with few preoccupations traceable to the Quetelet *mystique*, the contemporaries of Neyman and Pearson could pardonably regard it as a new refinement of an old technique, and indeed as a programme for promoting better and brighter significance test procedures.

Another reason for belated appreciation of the trail blazed by the Neyman-Pearson partnership emerges from a new use Wald—acclaimed by many as America's leading theoretical statis-



tician of our own generation\*—found for the comprehensive decision test procedure. Though the logical credentials of the comprehensive decision test must exert a compelling appeal to any student who approaches the issues involved with a fresh mind, it has a limited usefulness for a reason not yet discussed. In our fruitfly model set-up, we can reasonably limit our admissible hypotheses if appropriately fortified with background knowledge of the culture; but comparable situations are rare in comparison with those which the traditionalist significance test prescription claimed as its province. For instance, the general pattern of a hypothesis we might deem to be appropriate to the clinical trial is  $M_d \geq k$  and  $k$  may conceivably have any value in the range  $\pm 1$ , if we have no prior knowledge of the parameter ( $p_a$  above) definitive of our yardstick treatment group (A).

In such a situation as that of the Yule-Greenwood trial, we can indeed specify two exclusive alternatives in terms of a target value  $k$  as  $M_d < k$  and  $M_d \geq k$ ; but we have then seen that the power of the test—even if the test is a U.M.P. test—may be negligible for samples of any obtainable size. More generally, we may formulate this dilemma as follows. We suppose: (a) that the only admissible alternative to the null hypothesis  $p = p_0$  is  $p \geq p_h$ ; (b) that the rejection criterion assigns a probability  $\alpha$  to wrong rejection of the null hypothesis and  $\beta$  to wrong rejection of the particular alternative  $p = p_h$ . If the difference between  $p_h$  and  $p_0$  is infinitesimal,  $\beta \simeq 1 - \alpha$  (see Fig. 1) for any sample of finite size, and the uncertainty safeguard of the test procedure will be  $P_f \leq \beta$  if  $\alpha$  itself conforms to any acceptable criterion of conditional risk, e.g.  $P_f \leq 0.95$  if  $\alpha = 0.05$ . Thus prescription of sample size to ensure any acceptable uncertainty safeguard for decision test procedure is possible only if there is a finite difference between  $p_0$  and  $p_h$ .

Wald first pointed out the foregoing implications of the attempt to discriminate between exclusive alternative hypotheses consistent with a continuous range of admissible values of  $p$  from  $p_0$  to  $p_h$ , and proffered a recipe for dealing with a limited class of situations which then arise. If we may legiti-

\* *Journ. Amer. Stat. Ass.*, Vol. 46, 1951, pp. 242-4.

mately relinquish the attempt to discriminate between *exclusive* alternatives, we may sidestep this disability as follows. We assume two hypotheses  $H_a$  and  $H_b$  referable to definitive parameters  $M_a$  and  $M_b = (M_a + k)$ . If  $k$  is positive, our comprehensive test prescription is then as follows. We first define a rejection score  $x_r$  and sample size  $r$  to ensure acceptable values of  $\alpha$  and  $\beta$  such that

$$P_a(x \geq x_r) = \alpha \quad \text{and} \quad P_b(x < x_r) = \beta$$

For any hypothesis  $H_i$  that  $M_i < M_a$ , and for any hypothesis  $H_j$  that  $M_i > M_b$  we may then say:

$$P_i(x \geq x_r) < \alpha \quad \text{and} \quad P_j(x < x_r) < \beta$$

We can then assign to a specifiable rule an uncertainty safeguard  $\alpha \leq P_f \leq \beta$  if  $\beta > \alpha$ ,  $\alpha \geq P_f \geq \beta$  if  $\beta < \alpha$  and  $P \leq \alpha$  if  $\alpha = \beta$ , i.e.  $P_f \leq 0.05$  if  $\alpha = 0.05 = \beta$ . The rule is as follows:

- (i) If  $x \geq x_r$ , reject  $H_a$ , i.e. say that  $M > M_a$
- (ii) If  $x < x_r$ , reject  $H_b$ , i.e. say that  $M < M_b$

In the context of the clinical trial, we might then choose as our hypotheses  $M_a = 0.05$  and  $M_b = 0.10$ . In that event the outcome of the test would be to make either of the two following terminal statements:

- (a) the advantage of treatment B is greater than 5 per cent.
- (b) the advantage of treatment A is less than 10 per cent.

Now the statement that the advantage is less than 10 per cent does not exclude the possibility that it is also less than 5 per cent. Consequently, the test prescription does not provide a formula for a situation in which the only type of terminal statement consistent with the operational intention must take the form  $M \geq k$ . What it can do is to provide a formula for the commercial situation in which we assume: (i) that the consumer will tolerate a proportion of defective articles in excess of the producer's guarantee; (ii) the producer wishes to insure both against the risk of losing the consumer's good

will and against the risk of discarding consignments up to guaranteed standard.

Evidently a dilemma of this sort involves an ethical issue; and it is difficult to conceive situations confronting the research worker with a comparable choice which is also consistent with the ethic of science. If so, we may gratefully record the fact that the American way of life has provided a milieu in which it has been possible to explore fully the implications of the theory of the decision test. In so far as the undertaking has any bearing on its role as an instrument of statistical inference in scientific investigation, the outcome is that:

(a) any test procedure with an assignable and acceptable upper limit of uncertainty and the prospect of arriving at a decision on most occasions when applied is one which we can hope to prescribe for a very limited class of situations;

(b) no test procedure which conforms to both requirements last stated is relevant to the class of situations which the test batteries of the Analysis of Variance and the Analysis of Covariance claim as their province.

In this context, it is pertinent to state a third conclusion which will forestall the criticism that later chapters contain no detailed treatment of two such now fashionable procedures as last named. If we regard both requirements above stated as prerequisite to a consistently behaviourist approach to the usefulness of statistical procedures, neither of them calls for further comment in subsequent discussion of the terms of reference of a Calculus of Judgments. Indeed, we shall see (p. 420) that they rely on sampling distributions which fail to satisfy an additional and compelling requirement, not as yet explicitly stated.

## CHAPTER SIXTEEN

### INDUCTION AND DESIGN STOCHASTIC AND NON-STOCHASTIC

IN THE EXPOSITION of his own views on *The Design of Experiments*, R. A. Fisher declares that "the statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation." Here at least, we have grounds for agreement. We cannot indeed do justice to the claims of statistical inference if we refrain from asking: what is the role of any procedure referable to testing hypotheses as an instrument of scientific reasoning? This question, which we shall now scrutinise in conformity with Fisher's admirable counsel, imperatively raises others. What do we imply when we speak of the scientific method as inductive? What do we imply when we distinguish between *deduction* and *induction*?

Off-guard in a local brains trust with only a few seconds in which to frame a suitable reply, most scientific workers would answer the last question by defining: (a) deduction as reasoning from an assumed or known cause or set of axioms to an undiscerned effect or conclusion; (b) induction as reasoning from a known effect to an undiscerned cause. Alternatively, he or she might say in accord with the Concise Oxford Dictionary that: deduction is "inference from general to particular" and induction is "inferring of general law from particular instances." Either way, we are prone to regard induction as a reversal of the logical processes subsumed by the term deduction, whence by an all too common exercise of metonymy we identify it with the Backward Look.

The Oxford Dictionary definition is consistent with Mill's usage when he declared that "induction is inferring a proposition from propositions less general than itself." Professional logicians of a later vintage have been less preoccupied with the inadequacy of Mill's exposition of the process of scientific investigation than with the laxity of his concept of *cause* and with the irrelevance of the syllogistic formula to deduc-

tions incident to the interpretation of data. They have therefore done little to clarify the reorientation signalled by Neyman's use of the expression *inductive behaviour* and by what we have called the Forward Look in foregoing chapters. A brief reference to Mill's views is therefore pertinent to our theme. As is true of other philosophers of a past generation, it is easy to find in his writings anticipations of modern views. In one context, Mill defines induction unexceptionably but uninformatively as: "the operation of discovering and proving general propositions"; and we can concur readily with his view that a prerequisite to understanding the operation is "sufficient acquaintance with the process by which science has actually succeeded in establishing general truth." None the less, we must be wary of assuming either that the practice of scientific enquiry in his own time could provide the basis for such sufficient acquaintance or that a now acceptable interpretation of the term *general truth* as applied to a hypothesis is entirely consistent with the usage of Mill's contemporaries.

Mill's insistence on the unity of the scientific method had a wide emotive appeal to a contemporary popular front of secular opposition to ecclesiastical control of education; and the meaning he attaches to a general truth is wholly consistent with his message. There was as yet little provision for scientific research as a profession, and public pronouncements of the leaders of science, eloquent in the defence of unfettered curiosity and of conscientious fact-finding, disclose few agenda for a programme of enquiry on a level of discussion now likely to enlighten or to enlist the interest of the professional logician. For two reasons, the statement of a single formula for the scientific method embracing Biblical criticism, the study of history and social institutions, the theory of organic evolution, Newton's laws of motion and the discovery of lately unknown electromagnetic phenomena was indeed an easier undertaking for Mill and for his immediate following than it can now be for us. In the more mature sciences, hit-and-miss methods of the brilliant amateur exploring unknown territories in the domain of electricity or of animal behaviour have increasingly made way for a professional regimen of planned experimentation. Meanwhile, we think of the framework of interpretation

less as a photograph of nature in glorious technicolour than as an expanding code of recipes for human action. What endures and expands, as the building called science gains in breadth and stature, is not the scaffolding of metaphor. It is man's command over nature.

To make a just appraisal of Mill's viewpoint with due regard both to his immense erudition and to his indisputable sagacity, we must judge him in his own context, equally that of Landseer, of the traveller naturalists and of the Great Exhibition. In Mill's day and generation, the current view of the wonders of science was a Royal Academy view, an unfinished but ever unfolding landscape picture of a universe into which man is a recent, somewhat pitiable and becomingly apologetic intruder. Swinburne was nearer to the temper of our own time when he ended his memorable hymn with the lines "Glory to Man in the Highest, for Man is the master of things." So indeed was Bacon, when he proclaimed that the roads to human power and to human knowledge lie close together.

More than two centuries before Mill, the author of the *Novum Organum* had indeed pleaded with unsurpassed eloquence for an operational approach to the terms of reference of scientific enquiry *en rapport* with the reorientation of our own time; but the experimental method was then on trial. Like Bacon, Mill exalted its merits, but the essence of the experimental method, as he expounds it, is active interference with nature. His canons for the interpretation of the outcome of such interference scarcely transgress the boundaries of common sense, and shed no new light on the terms of reference of the undertaking. In so far as he had any prevision of what we now call the *Design of Experiments*, his main concern was to point out the dangers of reliance on what he called the *Hypothetical Method* in the search for a unique and final specification of cause and effect. He might well have had a less discouraging view of its usefulness if he had also had less inclination to announce a formula to fit all methods of investigation worthy of respect.

In what follows I shall use some terms introduced by Wrighton\* in a recent and thought-provoking statement

\* *Acta Genetica et Statistica Medica*, 1953.

of a novel approach to the scope of statistical inference. Accordingly, we shall distinguish between experiments of two kinds as *exploratory* and *holonomic*. In the latter we proceed from a limited set of hypotheses to deduce consequences which are factually verifiable and select a hypothesis or sub-set of hypotheses uniquely consistent with observation. We might speak of this as Mill's hypothetical method; but we shall not expose ourselves to his legitimate objections to its misuse if we undertake it with no illusions concerning our obligation to furnish in advance reasons for the choice of a uniquely admissible set of hypotheses. We shall also be able to appreciate why Mill's own generation had not "sufficient acquaintance with the process by which science has succeeded in establishing general truth," if we first explore its use in a field of enquiries which prompted Mill himself to express a very conservative—and, as now appears, factually unjustifiable—opinion about the usefulness of the experimental method.

The writer here deliberately selects an investigation into animal behaviour carried out under his own direction, because it will illustrate how the biologist himself approaches the problem of design when his aim is to solve a biological problem rather than to collect data suitable for demonstrating computations invoked by a significance test procedure. Conceding to Mill that our definition of the admissible set must itself have a factual basis, we assume that flies of the species *Drosophila melanogaster* have *hygro-receptors*, i.e. sense organs by which they recognise moisture, and *chemo-receptors*, i.e. sense organs by which they recognise specific chemical compounds. Our factual basis for this assumption is that normal flies of the species released in a closed space with choice of entering: (i) a dry or a moist, but otherwise similar, chamber congregate in the latter (positive *h*-response); (ii) a chamber containing dilute acetic acid and an otherwise similar chamber containing pure water congregate in the former (positive *c*-response). Let us also concede to Mill a second factual prerequisite to the design proposed, viz. that casual observation on the movements of the antennae of flies in search of food or of a sufficiently moist environment suggests the localisation of such receptors therein. Against this background information we may postulate a

uniquely admissible set of axioms or hypotheses embracing *all* factual possibilities :

- H.1 Both hygro-receptors and chemo-receptors of *D. melanogaster* are *exclusively* located in the antennae.
- H.2 The chemo-receptors are *exclusively* located in the antennae, but the hygro-receptors are not.
- H.3 The hygro-receptors are located *exclusively* in the antennae, but the chemo-receptors are not.
- H.4 Receptors of *neither* sort are exclusively (if at all) located in the antennae.

If we appropriately define our criterion of *negative* response, i.e. partition of released flies in approximately equal numbers in alternative choice-chambers, we may also classify *all* possible observations on a choice-chamber set-up involving flies of both sorts thus :

- O.1 normal flies give both a positive *h*-response and a positive *c*-response, but antennaless flies give both a negative *h*-response and a negative *c*-response ;
- O.2 normal flies as before, antennaless giving a positive *h*-response and a negative *c*-response ;
- O.3 normal flies as before, antennaless being *c*-positive and *h*-negative ;
- O.4 both normal and antennaless *c*-positive and *h*-positive.

Having stated both a comprehensive set of hypotheses and a comprehensive set of observations, we may now make the following deductions :

(i) If H.1 is true, normal flies will give both a positive *h*-response and a positive *c*-response, but antennaless flies will give both a negative *h*-response and a negative *c*-response.

(ii) If H.2 is true, normal flies will give both positive responses, but antennaless flies will give a positive *h*-response and a negative *c*-response.

(iii) If H.3 is true, normal flies will give both positive responses, but antennaless will give a positive *c*-response and a negative *h*-response.

(iv) If H.4 is true, normal flies will give both positive responses and antennaless will give both positive responses.



Up to this point, it is immaterial whether we speak of H.1–H.4 as axioms or hypotheses. Our next step is possible because each deduction (i)–(iv) is uniquely endorsed by an actual observation of a unique event or fact, if we can obtain by surgical or other procedure flies deemed to be healthy and comparable in all particulars other than lack of the antennae. Accordingly, we may exhibit a set of all conceivable observations and corresponding admissible hypotheses as below in a table with a + sign to indicate which observation is alone consistent with each hypothesis, and if recorded as such endorses one of four *terminal statements* T.1–T.4 consistent with our deductions (i)–(iv) above:

		<i>Observations</i>			
		O.1	O.2	O.3	O.4
<i>Hypotheses</i>	H.1	+			
	H.2		+		
	H.3			+	
	H.4				+

		<i>Terminal Statements</i>
	—	T.4 : if we observe O.4, we shall state that H.4 is true.
	—	T.3 : if we observe O.3, we shall state that H.3 is true.
	—	T.2 : if we observe O.2, we shall state that H.2 is true.
	—	T.1 : if we observe O.1, we shall state that H.1 is true.

The reader may here note the deliberate use of the future auxiliary in the formula of the terminal statement, emphasising the Forward Look which anticipates the final step called

*verification*. Subsequently, we may agree to drop it for brevity, when clear about what we are doing. The final step of *verification* itself is: (a) to release flies of both sorts with appropriate choice chambers for detection of the *h*-response only and flies of both sorts with appropriate choice chambers for detection of the *c*-response only; (b) to count the flies of each sort in each sort of choice chamber after a suitable lapse or—better still—at regular intervals.

In a perfectly designed experiment of this type, we shall need to forestall a legitimate objection of Mill to the hypothetical method by assuring ourselves that the *experimental* flies are in all relevant particulars like the *controls* except in so far as they lack antennae. Begg and Hogben (*Proc. Roy. Soc. B*, 133, 1946) sidestepped objections referable to shock, etc., due to surgical procedure by using flies of the mutant stock *antennaless*. With suitable diet such flies will reach maturity with two normal antennae, with one normal antenna or with no antennae at all. In the complete hypothesis and observation table we shall then have to allow for the inferred responses of four sorts of flies. Unless genotype *per se* introduces a new relevant variable, three of these should behave in the same way, viz. normal of normal stock, normal of antennaless stock and unilateral antennaless. The bilateral antennaless flies alone should give the responses specified as antennaless by (i)–(iv) above. The introduction of the controls is relevant to the process of verification only in so far as it is a check on the adequacy of the choice chamber set-up.

If we now speak of the comprehensive set of terminal statements endorsed by the experiment as a *rule of non-stochastic induction*, we do so because an investigator with high standards of experimental procedure would not be content with a specification of response as positive unless the overwhelming proportion of flies of a particular sort congregated in one alternative choice chamber, or as negative unless the proportions in each were approximately identical. Reliance on statistical tests to validate the interpretation of the outcome would indeed signify either or both of two defects of design: (a) lack of relevant genetical background information implicit in the claim that the final assertion (Wrighton's *terminal state-*

ment) embraces flies of the species as a whole; (b) lack of a satisfactory response criterion for the type of receptivity under investigation. In any event, our procedure is throughout consistent with the Forward Look. *Our induction is anticipatory deduction subject to the discipline of ineluctable fact.*

Before we now ask what is indeed characteristic of stochastic in contradistinction to non-stochastic induction, we may here usefully note that we can speak of interference with nature at more than one level. In many experiments of this sort we might legitimately adopt a surgical procedure instead of taking advantage of the fact that nature has provided us with two sorts of flies. Here therefore we interfere at the level of nurture alone. In studies on human behaviour we may take advantage of the fact that society arranges for the adoption of children; and we can set out reactions of fraternal or of identical co-twins brought up in the same or in different families in a hypothesis-observation table derivable without recourse to active interference at either the level of nature or the level of nurture. This suggests a possible programme of enquiry which may eventually bring within the same framework of interpretation disciplines whose conduct is consistent with the Forward Look and some disciplines in which we take the necessity of the Backward Look for granted. Be that as it may, the common disposition to assume that scientific enquiry on human beings in circumstances which exclude the active interference Mill so lightly regards as the hall-mark of the experimental method is essentially statistical—and less precise on that account—rests on a misunderstanding of what is an essential characteristic of designed experimentation.

Having defined a *holonomic* experiment as above, it is scarcely necessary to define the alternative category, except to say that some of the most noteworthy discoveries of the past—and this is largely true of discoveries in the domain of electricity, magnetism or physiology in Mill's time—were the outcome of experimentation undertaken to enlarge our factual knowledge rather than to interpret facts already known in a new way. The discovery of radioactivity and of radium should make us hesitate to dismiss the usefulness of such *exploratory* experimentation or to regard it as a thing of the past; and

indeed it is difficult to believe that designed enquiry can lead to the discovery of previously unsuspected natural phenomena except in so far as the investigator is alert to events which do not appear as entries in the hypothesis-observation table of the holonomic experiment. What makes experimentation here called holonomic of special interest is not that it is necessarily destined to supersede the alternative. It is instructive in this context because it has a special relevance to an intelligible distinction between non-stochastic and stochastic induction.

To get the distinction into sharper focus it is useful to divide experiments of the holonomic type into two categories, those which *may* and those which *must* lead to a conclusive outcome. In terms of design, Wrighton speaks of them respectively as *a posteriori* adequate and *a priori* adequate. Our last example was *a priori* adequate in the sense that there is a single positive sign in each column of the hypothesis-observation table, whence one and the same observation endorses only one of the set of terminal statements which justify the design. One example will suffice to specify the alternative class. We first predicate the following as background information. In fowls:

(a) one dominant gene substitution is responsible for the difference between the pea comb of the Partridge Cochins and the single comb of the Mediterranean breeds (leghorns, anconas, etc.);

(b) one dominant gene substitution is responsible for the difference between the rose comb of the Wyandottes or Hamburgs and the single comb;

(c) the interaction of both dominant genes in a gene complex otherwise like that of the single comb breeds leads to the development of the walnut comb of the Malay breeds. Accordingly, we may classify genotypes w.r.t. the four sorts of combs mentioned as:

- (i) *Single* pp.rr
- (ii) *Pea* PP.rr or Pp.rr
- (iii) *Rose* pp.RR or pp.Rr
- (iv) *Walnut* PP.RR; Pp.RR; PP.Rr; Pp.Rr

Thus we can set out the possible types to which we may assign a *single* offspring of all possible matings between two fowls *each* with the walnut comb in a hypothesis-observation grid (H.O.G.) as below:

	Hypotheses (Matings are:)	Observations (Offspring are):			
		O.1 Walnut	O.2 Pea	O.3 Rose	O.4 Single
1.	PP.RR $\times$ PP.RR	+			
2.	PP.RR $\times$ PP.Rr	+			
3.	PP.RR $\times$ Pp.RR	+			
4.	PP.RR $\times$ Pp.Rr	+			
5.	PP.Rr $\times$ Pp.RR	+			
6.	PP.Rr $\times$ PP.Rr	+	+		
7.	PP.Rr $\times$ Pp.Rr	+	+		
8.	Pp.RR $\times$ Pp.RR	+		+	
9.	Pp.RR $\times$ Pp.Rr	+		+	
10.	Pp.Rr $\times$ Pp.Rr	+	+	+	+

We may speak of each one of the comprehensive set of 10 hypotheses here disclosed as an *elementary* hypothesis. We make our table more compact by combining nine of them in three *composite* hypotheses which with the last in the foregoing make up the following 4-fold comprehensive set:

H.1. The parental pair is any one of the following:  
 PP.RR  $\times$  PP.RR; PP.RR  $\times$  Pp.RR; PP.RR  $\times$  PP.Rr;  
 PP.RR  $\times$  Pp.Rr; PP.Rr  $\times$  Pp.RR.

H.2. The parental pair is *either*:  $PP.Rr \times PP.Rr$  or  $PP.Rr \times Pp.Rr$ .

H.3. The parental pair is *either*:  $Pp.RR \times Pp.RR$  or  $Pp.RR \times Pp.Rr$ .

H.4. The parental pair is:  $Pp.Rr \times Pp.Rr$ .

Our H.O.G. of hypotheses referable to parental pairs of fowls with the walnut comb (regardless of sex) and of observations referable to one *individual* offspring, then assumes the form:

	O.1 (W)	O.2 (P)	O.3 (R)	O.4 (S)
H.1	+			
H.2	+	+		
H.3	+		+	
H.4	+	+	+	+

—  $T_4$  if O.4 say that H.4 is true.

—  $T_3$  if O.3 say that either H<sub>3</sub> or H<sub>4</sub> is true.

—  $T_2$  if O.2 say that either H<sub>2</sub> or H<sub>4</sub> is true.

—  $T_1$  if O.1 say nothing at all.

In this table only one column O.4 (S) contains one cell marked with the positive sign; and hence only one observation is consistent with a single hypothesis. Should it happen that only one chick comes to maturity, we can thus justify a terminal statement in favour of a single hypothesis of the 4-fold comprehensive set only if the chick also happens to have the single comb. The result then justifies the design; but no other result

could do so, if the intention is a terminal statement asserting that one of the four hypotheses is correct. In such a situation, Wrighton speaks of the design of the experiment as *a posteriori* adequate for the observation O.4.

*En passant* we may here note that our legitimate scope of inference is exactly the same if we look at the issue *retrospectively*, i.e. if asked to make a terminal statement of the same sort on the basis of information about two fowls each with the walnut comb. Though we are reasoning about a *past* event, our reasoning follows the path we traverse with eyes *forward* when we design an experiment to test the same set of hypotheses; and this may give us a clue to an acceptable answer to the question: how, if at all, is it possible to find a formula to embrace both acceptable reasoning about past events beyond the range of human interference and acceptable reasoning about situations we can control?

In the preceding discussion we have assumed that we can get information about only one offspring of the cross between two fowls with walnut combs classified in four possible categories as defined above. If we conceive the possibility of obtaining a large enough number of offspring to ensure representation of all possible phenotypes consistent with the parental mating class, we can classify conceivably observations on progeny of a single mating comprehensively as:

- O.1 walnut only (W)
- O.2 both walnut and pea but no other (W + P)
- O.3 both walnut and rose but no other (W + R)
- O.4 all four phenotypes (W + P + R + S)

On the assumption stated, it would thus be possible to make, as on page 382, the H.O.G. of an *a priori* adequate design in terms of the foregoing 4-fold comprehensive set of hypotheses.

We might speak of such a schema as *asymptotically* adequate *a priori* since there is a finite probability that the experiment will *not* fulfil our expectations, if *r* itself is finite as indeed it must be. That this is so, makes it a model of an alternative procedure of formulating a set of terminal statements in a

rule of stochastic induction. Accordingly, we shall refer to it henceforth as *Model I*.

*Model I.* Since the number ( $r$ ) of offspring two fowls with the walnut comb may produce is finite, each hypothesis is consistent with the occurrence of fraternities exclusively made up of W alone; but such occurrences will be rare when  $r$  is large unless H.1 is true. If H.2 is true, we may expect to meet fraternities of phenotypes W alone, P alone or W and P

	O.1	O.2	O.3	O.4
	(W)	(W + P)	(W + R)	(W + P + R + S)
H.1	+			
H.2		+		
H.3			+	
H.4				+

$T_4$  if O.4 assert H.4 is true.  
 $T_3$  if O.3 assert H.3 is true.  
 $T_2$  if O.2 assert H.2 is true.  
 $T_1$  if O.1 assert H.1 is true.

together, the last most often. If H.3 is true, we may expect to meet W alone, R alone or W and R together, the last most often. If H.4 is true, we may expect to meet fraternities of W, P, R or S alone, any pair (W + P, W + R, W + S, P + R, P + S, R + S) of the four phenotypes alone, any three alone (W + P + R, W + P + S, W + R + S, P + R + S) or all four (W + P + R + S). With due regard to what we shall indeed most commonly meet in the totality of our experience, we are free to classify all possible observations



on an  $r$ -fold fraternity with parents both of phenotype W as containing:

O.1 W only.

O.2 At least one P but no R and no S.

O.3 At least one R but no P and no S.

O.4 Either at least one S or at least one of *both* R and P.

The probability assigned by the Theory of the Gene to the four possible phenotypes to which an individual offspring of each class of matings is itself assignable are

	W	P	R	S
H.1	1	0	0	0
H.2	$\frac{3}{4}$	$\frac{1}{4}$	0	0
H.3	$\frac{3}{4}$	0	$\frac{1}{4}$	0
H.4	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$

Here we note that each of the four hypotheses, heretofore specified in *qualitative* terms, of our comprehensive set is also distinguishable by a numerical specification, viz. the three (one being redundant) parameters definitive of the randomwise sampling distribution of phenotypes among the offspring of the relevant mating. Only when it is possible to identify each hypothesis of the comprehensive set uniquely by a parameter or parameters definitive of a unique sampling distribution can we take the first step from the domain of a rule of non-

	O.1	O.2	O.3	O.4
	(W only)	(at least one P, but no R and no S)	(at least one R, but no P and no S)	At least one S or at least one R and at least one P also
H.1	1	0	0	0
H.2	$(\frac{3}{4})^r$	$1 - (\frac{3}{4})^r$	0	0
H.3	$(\frac{3}{4})^r$	0	$1 - (\frac{3}{4})^r$	0
H.4	$(\frac{9}{16})^r$	$(\frac{3}{4})^r - (\frac{9}{16})^r$	$(\frac{3}{4})^r - (\frac{9}{16})^r$	$1 - 2(\frac{3}{4})^r + (\frac{9}{16})^r$

stochastic to a rule of stochastic induction. We may then take the next step. We lay out an H.O.G. (leaving blank the specification of the terminal statements) with probability entries in place of the  $+$  sign for concurrence of hypothesis and observation. From the parameters shown above we easily derive the H.O.G. shown at foot of page 383.

This design is not *a priori* adequate as it stands, since  $r$  is finite; but we can imagine what it would imply if  $r$  became indefinitely large, so that each diagonal cell would contain a unit and every other cell a zero. It would then look like the table (p. 382) of an *a priori* adequate set-up; and we should be able to operate the rule of induction which subsumes the exhaustive 4-fold set of terminal statements.

Let us now imagine that  $r$  is so large that each diagonal entry is numerically well above 0.5. We may then say that

- T.1 if all the offspring are W, we *shall* assert that H.1 is true;
- T.2 if the offspring include at least one P but no R or S, we *shall* assert that H.2 is true;
- T.3 if the offspring include at least one R but no P or S, we *shall* assert that H.3 is true;
- T.4 if the offspring either include at least one S or at least one R *and* at least one P also, we *shall* assert that H.4 is true.

More than 50 per cent of the observations we shall meet in a long enough sequence of trials will be specifiable as:

- O.1 if H.1 is true
- O.2 if H.2 is true
- O.3 if H.3 is true
- O.4 if H.4 is true

If we consistently follow the rule last stated, we shall say that H.1 is true only if we make the observation O.1 and this we shall do in more than 50 per cent of the situations we shall encounter if H.1 is indeed true. In more than 50 per cent of the situations we shall encounter when H.1 is true the assertion we shall make will be true, i.e. more than 50 per cent

of the assertions we make will be true, if we adhere to the rule regardless of the outcome. Similar remarks apply *mutatis mutandis* to H.2, H.3 and H.4. *Whichever* hypothesis is true, more than 50 per cent of our assertions will therefore be true, if we follow the rule consistently. At the risk of being wrong sometimes, we can thus frame a rule which will guarantee both that we: (a) make a decisive judgment; (b) do so correctly more often than otherwise.

To do so we have: (a) to agree about what risk of erroneous statement we are prepared to take; (b) to fix accordingly the size ( $r$ ) of fraternities w.r.t. which we record the relevant observations. First, we note that the lowest probability cited as a diagonal cell entry refers to an observation which will *least* often lead us to identify the correct hypothesis correctly. In the last table this is the entry on the right at the foot. If we denote by  $P_t$  the proportion of our assertions which will be right in the long run, we can therefore write:

$$P_t \geq 1 - 2\left(\frac{3}{4}\right)^r + \left(\frac{9}{16}\right)^r$$

For the values  $r = 12, 13, 14$ , we may thus tabulate the relevant expressions thus

$r$	$1 - \left(\frac{3}{4}\right)^r$	$1 - 2\left(\frac{3}{4}\right)^r + \left(\frac{9}{16}\right)^r$
12	0.969	0.937
13	0.976	0.953
14	0.982	0.965

To make the probability of correct assertion at least 95 per cent, we must thus confine our attention to fraternities of  $r \geq 14$ . If we denote the risk of erroneous statement by  $P_f = (1 - P_t)$  we may say that  $P_f < 0.05$  if  $r \geq 14$  for the *rule of stochastic induction* subsumed under the foregoing statements. We have elsewhere spoken of this risk as the *uncertainty safeguard* of the rule. We now see more explicitly that it is *a property of the rule in its entirety*. It is not the probability that we shall be right, if we restrict our verdicts to observations deemed to endorse any individual terminal statement. If H.4 is true, we shall always be wrong, if we assert H.1 is true when we observe O.1; and all we can legitimately say about the probability that the

*individual* terminal statement will be true is that it may be either zero or unity.

In one respect, the Model I design is highly artificial. We have approached the distinction between a stochastic and a non-stochastic rule of induction by tracing the steps traversed in a *designed* experiment, but we have not asked ourselves: what end in view has the design? We have provisionally distinguished between a terminal statement which is the assertion that a hypothesis is actually true and a hypothesis which is the assertion of a statement conceivably true in the absence of evidence to the contrary; and we have provisionally defined *a priori adequacy* of design in terms of one to one correspondence between terminal statements and hypotheses. We have also quite arbitrarily exercised the liberty of combining elementary in composite hypotheses to ensure the possibility of *a priori* adequacy; but such liberty is inconsistent with the notion of design unless the comprehensive set of terminal statements is *acceptable* in the sense that it supplies answers consistent with a presumptive operational intent. More explicitly therefore, we should speak of an experiment as *a priori* adequate only if there is one to one correspondence between the members of the comprehensive set of hypotheses—composite or otherwise—and a comprehensive set of *acceptable terminal statements*.

If we then ask whether the definition of a 4-fold comprehensive set of hypotheses and of terminal statements adopted in the foregoing model situation is consistent with an intelligible criterion of acceptability, the answer must be *no*. A practical poultry breeder having read so far might with justice protest that we have butchered the presumptive operational intent of the experiment to indulge in a statistical holiday. We have defined hypothesis IV in terms of a unique mating which specifies uniquely the genotype of each parent regardless of sex, but hypotheses I–III each embrace more than one mating, only one such mating (PP.RR  $\times$  PP.RR) being unique w.r.t. the sex of the parents. If we had specified all the ten different hypotheses each referable to a unique mating on the assumption that the operational end in view is to identify a unique parental *pair* of genotypes or all the sixteen different hypotheses each

referable to a unique mating if our aim is to identify the genotype of cock and hen *separately*, no classification of possible observations based on a mating of the cock and the hen *inter se* would have been consistent with an *a priori* adequate design. We have indeed invoked a 4-fold comprehensive set of hypotheses, and hence of admissible terminal statements, to illustrate a theoretical nicety without asking whether it accomplishes any useful result in the realm of action.

The writer here concedes the deliberate commission for expository purposes of an error which is a serious, but by no means the only serious, ground for criticism of the unique null hypothesis drill for the therapeutic, prophylactic or agricultural field trial. The truth is that the experiment under discussion has no intelligible operational intent with which the design is consistent. If we really want to know the genotype of a cock and of a hen each with a walnut comb, we shall not mate them *inter se*. We shall mate each with a fowl of opposite sex and of the single comb type. In that event, and only so, our design will be *a priori* adequate to the presumptive operational intent.

In so far as the design of an experiment is consistent with the proper uses alike of stochastic or of non-stochastic induction, we may now set out the steps thus:

(i) to what *question* or questions do we seek an answer or answers?

(ii) what *background knowledge* of the situation in which we seek an answer or answers is initially available?

(iii) what opportunities of *relevant observation* does the proposed situation offer?

(iv) what form must the set of *terminal statements* have to be acceptable in terms of (i) and attainable in terms of (ii) and (iii)?

(v) what *comprehensive set of hypotheses* is consistent as a whole with (ii)?

(vi) what specification of the members of the comprehensive set of hypotheses is *a priori* adequate to the design in terms of (iii) and (iv)?

*Model II.* To illustrate the place of each of these in the design of an experiment, let us now examine a model set-up of the

sort we commonly invoke to illustrate the uses of the theory of probability, though we shall do so without actually transgressing the boundaries of non-stochastic induction.

1. To what question do we seek an answer?

The mean cash value ( $M$ ) of the tickets in an urn.

2. What background knowledge may we assume?

(a) the urn contains  $N$  tickets of cash value £  $x$ ,  $x + 1$ ,  $x + 2 \dots x + N - 1$ ,  $x$  being an integer;

(b) there are not less than three nor more than six tickets in the urn ( $3 \leq N \leq 6$ );

(c) the cash value of no ticket exceeds £7, i.e.  $(x + N - 1) \leq 7$ .

3. What observations shall we be free to make?

The mean cash value ( $m$ ) of two tickets removed simultaneously from the urn.

Before examining what terminal statements are acceptable in terms of the operational intent (1) and attainable in terms of our observations (3) interpreted against our background knowledge (2), let us set out as below the construction of the Hypothesis-observation table and the conclusions we can derive from it.

$N$	All possible sequences consistent with postulates						$m$ (in intervals of 0.5)	
	$M$							
6	2	3	4	5	6	7	4.5	2.5 — 6.5
5	2	3	4	5	6	.	4.0	2.5 — 5.5
	3	4	5	6	7	.	5.0	3.5 — 6.5
4	2	3	4	5	.	.	3.5	2.5 — 4.5
	3	4	5	6	.	.	4.5	3.5 — 5.5
	4	5	6	7	.	.	5.5	4.5 — 6.5
3	2	3	4	.	.	.	3.0	2.5 — 3.5
	3	4	5	.	.	.	4.0	3.5 — 4.5
	4	5	6	.	.	.	5.0	4.5 — 5.5
	5	6	7	.	.	.	6.0	5.5 — 6.5

*Hypothesis Observation Grid*

Hypothesis	Observations ( <i>m</i> )									
( <i>M</i> )	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	
3	+	+	+							
3.5	+	+	+	+	+					
4.0	+	+	+	+	+	+	+			
4.5	+	+	+	+	+	+	+	+	+	
5.0			+	+	+	+	+	+	+	
5.5					+	+	+	+	+	
6.0							+	+	+	

*Permissible Terminal Statements*

Observation ( <i>m</i> )	Lower Limit of <i>M</i>	Upper Limit of <i>M</i>
2.5	3.0	4.5
3.0	3.0	4.5
3.5	3.0	5.0
4.0	3.5	5.0
4.5	3.5	5.5
5.0	4.0	5.5
5.5	4.0	6.0
6.0	4.5	6.0
6.5	4.5	6.0

Evidently, we cannot give an exact answer to the question stated in 1 above, but if we are content to accept the most *precise* answer the data of any one experiment can ever endorse, our acceptable terminal statement will take the form:

$$M = m \pm 2$$

If we decline to pass judgment on the result of the experiment unless the score (*m*) is a whole number, we could always assert truthfully that

$$M = m \pm 1.5$$

We may thus say that the design of the experiment is *a priori* adequate if we deem a terminal statement of the form  $M = m \pm 2$  to be acceptable. If the end in view is to make the more precise statement  $M = m \pm 1.5$  the design of the experiment does not fulfil the requirement of *a priori* adequacy but it is *a posteriori* adequate for situations in which the observed score is an integer. In short, the best answer we can guarantee to give in all circumstances is a terminal statement which picks out the smallest number of individual hypotheses consistent with each possible observation we may make.

This criterion of choice has no special interest in the domain of Model I, because no numerical specification of our hypotheses is there meaningful *vis-à-vis* a rule of non-stochastic induction; and no hypothesis is specifiable for the purposes of stochastic induction by a single parameter. Here one number suffices to label an elementary hypothesis, and the comprehensive set of elementary hypotheses are meaningfully presentable as an ordered set within which we can make a more parsimonious disposition of composite hypotheses each definable as an *interval*. If the longest interval specified by any such composite hypothesis endorsed by any one of the corresponding set of terminal statements is consistent with the *precision* which defines our criterion of acceptability, the classification of our hypotheses satisfies the requirements of *a priori* adequacy.

Though we have introduced no stochastic considerations in our discussion of the foregoing model, we have here reached a conclusion which points the way to an extension of the terms of reference of a calculus of judgments consistent with the Forward Look, embracing the theory of the decision test as a special case and resolving the dilemma arising from its limited usefulness. A third set-up much like the last but simpler for our purpose will serve to bring into sharper focus both the line of demarcation between stochastic and non-stochastic induction and also to clarify some of the essential features of the type of stochastic induction subsumed by the term *interval estimation*.

*Model III.* For what follows we may briefly state the data relevant to the initial problem thus:



(a) *Background Knowledge*

An urn contains six tickets of cash value  $\pounds p, p + 1, p + 2 \dots p + 5$  on the understanding that  $p$  is an integer in the range  $1 \leq p \leq 6$ .

(b) *Permissible Observations*

We take two tickets successively, replacing the first and shaking up the urn thoroughly before taking the second and record the mean cash value ( $x_m$ ).

(c) *Required Information*

What is the least cash value ( $p$ ) of any ticket?

We shall approach the problem last stated both in terms of non-stochastic and of stochastic induction. A rule of stochastic induction must invoke a preliminary prescription of the sampling distribution referable to each hypothesis endorsed by a terminal statement. This we shall first do. For any value of  $p$ , we easily obtain the following random distribution of the mean score of 2-fold samples by recourse to a chessboard diagram:

*Distribution of 2-fold sample mean (with replacement)*

Frequency	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$
Sample score	$p$	$p + \frac{1}{2}$	$p + 1$	$p + 1\frac{1}{2}$	$p + 2$	$p + 2\frac{1}{2}$
Frequency	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	
Sample score	$p + 3$	$p + 3\frac{1}{2}$	$p + 4$	$p + 4\frac{1}{2}$	$p + 5$	

We may accordingly lay out a hypothesis-observation table for values of  $p$  in the range  $1 \geq p \geq 6$  as below; but we shall not signify concurrence of observation and hypothesis by a + sign. Instead we enter in each cell entitled to the sign of consent a number ( $>1$ ) which expresses as a multiple of  $6^{-2}$  the long-run frequency of each observation consistent with a particular hypothesis specifying a unique value of  $p$ ; and we may speak of all cells so labelled as the *region endorsed* by the background information. If we look at the experiment in the domain of non-stochastic induction, inspection of the table

then shows that the most general form of acceptable terminal statement consistent with the specification of requisite information (*c*) by the endorsed region of the H.O.G. is:

$$x_m \geq p \geq x_m - 5$$

This statement is *always* true and embodies as much as we can ever hope to say on the basis of the permitted class of observations, if we seek to frame a rule of decision with the intention implicit in the italicised adverb. Now different hypotheses (values of  $p$ ) here assign different long-run frequencies to one and the same observation ( $x_m$ ). From the foregoing set-up of the sample distribution we see that 34 ( $6^{-2}$ ) is the proportion of all observations in the range  $(p + 0.5) \leq x_m \leq (p + 4.5)$  for all values of  $p$ , i.e. we retain 94.4 per cent of the background information consistent with our long-run experience, if we exclude observations outside this range from the *endorsed region*. In that event, 94.4 per cent of our assertions will be correct for any value  $p$  may have, if we always assert that

$$x_m - 0.5 \geq p \geq x_m - 4.5$$

Thus we may narrow the interval in which we deem  $p$  to lie thereby achieving a higher level of terminal acceptability, i.e. a more precise answer to our question, if we are content to take an assignable risk that the rule will sometimes fail to give a true answer. If a stochastic procedure cannot guarantee a more precise answer than a rule of non-stochastic induction can endorse in the same set of situations, the risk we thus take is fruitless. We cannot then meaningfully speak of the statements it endorses as acceptable.

Before we travel further, let us scrutinise more closely what we *can* legitimately say and what we *cannot* legitimately say about the H.O.G. in the stochastic domain. We may set out the two rules of procedure in terms of the comprehensive set of possible observations we may make and the corresponding terminal statements subsumed by each rule as in the table on p. 394.

*Model III*  
(The cell entries are *relative frequencies*)

Hypotheses ( $p =$ )	Observations $x_m =$																					
	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	
1	(1)	2	3	4	5	6	5	4	3	2	(1)											
2			(1)	2	3	4	5	6	5	4	3	2	(1)									
3					(1)	2	3	4	5	6	5	4	3	2	(1)							
4						(1)	2	3	4	5	6	5	4	3	2	(1)						
5							(1)	2	3	4	5	6	5	4	3	2	(1)					
6									(1)	2	3	4	5	6	5	4	3	2	(1)			

## Model III

Observation (when $x_m$ is):	Terminal Statements (we shall assert that):	
	Stochastic ( $P_f = \frac{1}{18}$ )	Non-Stochastic
1.0	?	$p = 1$
1.5	$p = 1$	$p = 1$
2.0	$p = 1$	$1 \leq p \leq 2$
2.5	$1 \leq p \leq 2$	$1 \leq p \leq 2$
3.0	$1 \leq p \leq 2$	$1 \leq p \leq 3$
3.5	$1 \leq p \leq 3$	$1 \leq p \leq 3$
4.0	$1 \leq p \leq 3$	$1 \leq p \leq 4$
4.5	$1 \leq p \leq 4$	$1 \leq p \leq 4$
5.0	$1 \leq p \leq 4$	$1 \leq p \leq 5$
5.5	$1 \leq p \leq 5$	$1 \leq p \leq 5$
6.0	$2 \leq p \leq 5$	$1 \leq p \leq 6$
6.5	$2 \leq p \leq 6$	$2 \leq p \leq 6$
7.0	$3 \leq p \leq 6$	$2 \leq p \leq 6$
7.5	$3 \leq p \leq 6$	$3 \leq p \leq 6$
8.0	$4 \leq p \leq 6$	$3 \leq p \leq 6$
8.5	$4 \leq p \leq 6$	$4 \leq p \leq 6$
9.0	$5 \leq p \leq 6$	$4 \leq p \leq 6$
9.5	$5 \leq p \leq 6$	$5 \leq p \leq 6$
10.0	$p = 6$	$5 \leq p \leq 6$
10.5	$p = 6$	$p = 6$
11.0	?	$p = 6$

The queries in this table raise an issue we shall face once more in connexion with Models V (a) and V (b) (pp. 407-416). For the present, we shall disregard them. If we denote our hypotheses as  $H_n$  for the assertion  $p = n$ , etc., in the range  $2 \leq n \leq 5$ :

- (i) If  $H_n$  is correct every terminal statement we make in accordance with our rule will be correct except when  $x_m = p$  or when  $x_m = (p + 5)$ ;

(ii) situations in which  $x_m = p$  or  $x_m = (p + 5)$  make up  $\frac{1}{18} = P_f$ , of the totality of our experience within the assumed framework of unending repetition;

(iii) if therefore we make a terminal statement consonant with the rule whenever we take a sample regardless of its value, only  $\frac{1}{18}$  of our assertions will be false in the long run.

Now consider the case when  $n = 1$  or  $n = 6$ . As it stands, the table will entitle us to make no statement. If we refrain from doing so, the numbers in the corresponding rows specify  $\frac{3.5}{3.6}$  of our total experience, and we can err in our terminal statements

if  $H_1$  is true when  $x_m = 6.0$  with probability  $\frac{1}{3.5} < P_f$

if  $H_6$  is true when  $x_m = 11.0$  with probability  $\frac{1}{3.5} < P_f$

Alternatively, we may modify the rule by the addendum

Say  $p = 1$  if  $x_m = 1, 1.5$  or  $2.0$  with probability  $1.0$

Say  $p = 6$  if  $x_m = 10, 10.5$  or  $11.0$  with probability  $1.0$

Again, we shall err only if  $x_m = 6.0$  when  $p = 1$  and  $x_m = 6.0$  when  $p = 6$ , and

If  $H_1$  is true  $x_m = 6.0$  with probability  $\frac{1}{3.6} < P_f$

If  $H_6$  is true  $x_m = 11.0$  with probability  $\frac{1}{3.6} < P_f$

In either event, every admissible hypothesis prescribes that the proportion of false terminal statements will be equal to or less than  $P_f = \frac{1}{18}$ . Thus we can say that at least  $\frac{17}{18}$  of our assertions will be true in the long run, if we operate the rule consistently. This is what we mean by saying that the uncertainty safeguard  $P_f \leq \frac{1}{18}$  specifies the risk of error incurred by following the rule.

It is most important to have no illusions about what we rightly mean by following the rule *consistently* in this context. We are talking about the *totality* of our experience in the repetitive framework of randomwise sampling. We are not talking about a fraction of this totality such as the class of situations which arise when  $x_m = 4.0$ . Thus we cannot say that

the risk of false assertion is  $P_f \leq \frac{1}{18}$  when we make the particular assertion  $1 \leq p \leq 3$  if  $x_m = 4.0$ , implying thereby that the probability that such an assertion is true is  $P_t \geq \frac{17}{18}$ . Our table shows that  $x_m$  may indeed be 4.0 when  $p = 4$ , and we are exploring a range of admissible possibilities in which  $p$  actually has a particular value which may be  $p = 4$ . Should it happen that  $H_4$  is true, we should always be wrong in asserting that  $1 \leq p \leq 3$  when  $x_m = 4.0$ , and we may write  $P_{f,n} = 1$ , so that  $P_{t,n} = 0$  respectively for the *conditional* uncertainty safeguard and conditional stochastic credibility of the assertions embodied in our rule for all *particular* situations arising when  $x_m = n$  in the range  $1 \leq p \leq 5$ .

This distinction brings out a fundamental difference between what is meaningful, if we adopt the *Forward Look* which embraces the outcome of a rule conceived in behaviourist terms, and what we deem to be meaningful, if we adopt the *Backward Look* consistent with the location of probability "in the mind." Of the endorsed region of our table it is true to say w.r.t. any *conceivable* particular value of  $p$

$$(p + 0.5) \leq x_m \leq (p + 4.5) \quad \text{with probability } (1 - P_f) \geq \frac{17}{18} \quad (i)$$

We may express this symbolically as

$$P(p + 0.5 \leq m \leq p + 4.5) \geq \frac{17}{18} \quad . \quad (ii)$$

Of our table *as a whole*, we can also rightly say for all values of  $p$ :

$$(x_m - 0.5) \geq p \geq (x_m - 4.5) \quad . \quad . \quad (iii)$$

The formal identity of (i) and (iii) conceals a factual difference. Actually, we shall encounter in the totality of our experience any value of  $x_m$  consistent with the value of  $p$  we insert; but we can meet only one value of  $p$  in the same framework of repetition, since only one hypothesis of our admissible set is true of the assumed homogeneous universe (our urn) of choice. Thus we might speak of  $x_m$  as a factual variable and  $p$  as a conceptual variable. Of one we can make statements in the domain of events to the effect:  $x_m$  is . . . if  $p$  is . . . We have

left the domain of events if we say:  $p$  is . . . if  $x_m$  is . . . In the domain of events,  $p$  is a constant for whatsoever value of  $x_m$  we observe. If we interpret it retrospectively as a statement about the probability that  $p$  lies in a particular range when we already know that  $x_m$  has a particular value, we therefore retreat into the shadowy domains of probabilities in the mind. We do so explicitly if we embody (iii) in a statement analogous to (ii), viz.:

$$P(x_m - 0.5 \geq p \geq x_m - 4.5) \geq (1 - P_f)$$

In any meaningful sense consistent with the behaviourist outlook, the above refers only to the long-run frequency of correct assertion within the framework of a rule which we operate without knowing what value of  $x_m$  we may meet on any single occasion for its application. It follows that the design must be *a priori adequate* for the form of terminal statement deemed to be acceptable and this is inconsistent with a procedure which puts no restriction on what form of terminal statement is indeed acceptable. In the domain of *interval estimation* illustrated by our Model III, this implies a statement of the length of the interval we deem to be satisfactory. Only then can we define the size of sample consistent with an assigned upper limit of acceptable risk.

For Model III the procedure is as follows. Instead of assuming arbitrarily for expository purposes that we are content with a terminal statement to the effect that  $p$  lies in the range from  $(x_m - 4.5)$  to  $(x_m - 0.5)$  if we can assign an uncertainty safeguard  $P_f < 0.06$  to our rule, we shall now assume that no terminal statement will be acceptably precise unless our rule places  $p$  in the interval  $(x_m - 4)$  to  $(x_m - 1)$  with an uncertainty safeguard  $P_f \leq 0.05$ . Our first step is to tabulate distributions of the score mean  $(x_m)$  3-fold, 4-fold, 5-fold, etc., samples. This we can do by successive application of the chessboard device which we invoked to define the 2-fold sample mean score distribution on page 393. We may then make a more condensed table of probabilities as shown in the table on page 398.

To say that  $p$  lies in the range  $(x_m - 4)$  to  $(x_m - 1)$  inclusive is formally equivalent to the statement that  $x_m$  lies in the range

$(p + 1)$  to  $(p + 4)$ , if we interpret the identity as valid for the operation of the rule in its entirety. Our uncertainty safeguard to the former assertion is thus the probability that  $x_m$  will lie outside the range  $(p + 1)$  to  $(p + 4)$ ; and our table shows that this exceeds 0.05 unless  $r > 4$ . Thus we shall be able to design a procedure which accomplishes the end in view, only if we base our assertions on the outcome of taking with replacement

Probability that $x_m$ lies					
Outside the Range (inclusive)	For samples of				
	2	3	4	5	6
$p + 0.5$ to $p + 4.5$	0.056	0.037	0.008	0.005	0.001
$p + 1.0$ to $p + 4.0$	0.167	0.093	0.054	0.032	0.020
$p + 1.5$ to $p + 3.5$	0.333	0.324	0.194	0.196	0.122
$p + 2.0$ to $p + 3.0$	0.556	0.519	0.478	0.443	0.412

5-fold or larger samples randomwise from the urn. It is worthy of comment that we cannot here express an uncertainty safeguard by the identity  $P_f = \epsilon$  in terms of a preassigned level ( $\epsilon$ ) of acceptability. All we can say is  $P_f < \epsilon$ , or  $P_f \leq \epsilon$ . This is because we are dealing in this context realistically with a discrete distribution.  $P_f = \epsilon$  is justifiable only if we can justifiably invoke a continuous distribution of sample score values.



## CHAPTER SEVENTEEN

# RECIPE AND RULE IN STOCHASTIC INDUCTION

WITHIN WHAT WE may regard as the legitimate terms of reference of a calculus of judgments consonant with a behaviourist outlook, it has been customary to make a sharp distinction between *test* procedures and *interval*—in contradistinction to *point*—estimation. We have examined the credentials of two views of test procedure in Chapter Fifteen; and in current use the term interval estimation also subsumes divergent views propounded by the same opposing schools. The divergence between Neyman's *Theory of Confidence* and R. A. Fisher's doctrine of *Fiducial Probability* did not become apparent to many writers on statistical theory till many years after the appearance of the original publication of their views. This is partly because the arithmetical recipes prescribed by them are in many situations identical; but if we probe more deeply we may discern another source of misunderstanding. We have seen that the disputable issues involved in test procedure did not come sharply into focus till Wald expounded the views of Neyman and E. S. Pearson. If we now adopt the approach to stochastic induction in a publication (*op. cit.*) by Wrighton,\* we shall see that a comparable reorientation of the problem of interval estimation is overdue.

Several circumstances have conspired to retard such a reorientation. One of these is the disposition to emphasise the distinction between test procedure and interval estimation at the wrong level, as when we identify the former with the domain of situations admitting discretely bounded alternative hypotheses and the latter with that of situations consistent with a continuum of parameter values each conceptually definitive of an initially admissible elementary hypothesis. What is more important emerges in the distinction Wrighton draws between exploratory and holonomic experimentation. If we take seriously Neyman's own interpretation of a rule

\* *Acta Genetica et Statistica Medica*, 1953.

of stochastic induction as a rule stated in advance and consistently followed regardless of the outcome, we shall too readily overlook some of its more challenging implications unless we also ask what task the rule accomplishes. We must then impose a criterion of acceptability on the terminal statements subsumed by the rule. If we likewise concur in the full implications of Neyman's interpretation of the uncertainty safeguard as a risk associated with the operation of the rule *in its entirety*, we must also concede that: (a) each of a comprehensive set of *acceptable* terminal statements must endorse one, and only one, of a comprehensive set of hypotheses; (b) there must be a corresponding terminal statement to endorse each member of the comprehensive set of hypotheses.

Such is the principle of *a priori adequacy* which Wrighton postulates as an essential property of the stochastic hypothesis-observation grid, and his approach by way of the H.O.G. of a holonomic experiment in the non-stochastic domain suggests a sufficient reason for Neyman's failure to recognise its relevance to the Confidence controversy which is the topic of a later chapter. If we accustom ourselves to regard the H.O.G. as a visualisation of stochastic induction in the discrete domain of classical models such as Model III of the last chapter, it is easy to conceive each cell as an infinitesimal element of area and hence an effortless step to regard a graph exhibiting interval estimation in the continuum as a recipe—adequate or otherwise—for making a rule of procedure consistent with the Forward Look. Contrariwise, it is all too easy to overlook the elementary logic of the procedures we adopt, if we follow the historic path, *i.e.* if we start with the theory of interval estimation in the factually nebulous domain of an infinity of hypotheses w.r.t. each of which the relevant sampling distribution admits an infinity of score values.

Accordingly, we shall now explore a set of model situations which will disclose essential features of stochastic induction with a view to:

- (a) exhibiting test procedure and interval estimation as variants of a common pattern of reasoning;
- (b) making explicit certain canons prerequisite to the complete definition and prescription of a rule of stochastic induction.

# RECIPE AND RULE IN STOCHASTIC INDUCTION

*Models IV (a) and IV (b).* Consider the following model situations:

(a) Two pennies come from a bag containing pennies of 3 sorts:

- A minted with 2 heads
- B minted with 2 tails
- C normal

(b) Two pennies come from *one* of 2 bags, respectively containing:

- (i) normal pennies (C) and pennies with 2 heads (A);
- (ii) normal pennies (C) and pennies with 2 tails (B).

We shall suppose that the acceptable terminal statement of an experimental design is to identify the *pair* itself. In that event our comprehensive set of hypotheses need take no stock of order. For the Model IV (a) set-up, we may specify our elementary hypothesis in qualitative terms as: AA, AB, AC, CC, CB, BB. For the Model IV (b) set-up they are: AA, AC, CC, CB, BB. For either, we may specify all the relevant information for a stochastic design based on 4 tosses of each penny chosen as in Table I below:

TABLE I

Type	Actual No. of Heads in the pair	No. of Heads <i>observed</i> in the 4-fold joint toss								
		0	1	2	3	4	5	6	7	8
AA	4	0	0	0	0	0	0	0	0	1
AC	3	0	0	0	0	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$
AB	2	0	0	0	0	1	0	0	0	0
CC	2	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$
BC	1	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$	0	0	0	0
BB	0	1	0	0	0	0	0	0	0	0

# STATISTICAL THEORY

This method of specifying our hypotheses is valid for either model; and is alone adequate to label each distinctively in the Model IV (a) set-up; but it is not the only way in which we can do so in the Model IV (b) set-up, where we exclude AB in

TABLE II

## First Rule of Exclusion

Hypothesis		Observation (Heads)								
Type	No of Heads	0	1	2	3	4	5	6	7	8
AA	4									+
AC	3					+	+	+	+	-
CC	2	-	-	+	+	+	+	+	-	-
BC	1	-	+	+	+	+				
BB	0	+								

## Second Rule of Exclusion

Hypothesis		Observation (Heads)								
Type	No. of Heads	0	1	2	3	4	5	6	7	8
AA	4									+
AC	3					-	+	+	+	+
CC	2	-	-	+	+	+	+	+	-	-
BC	1	+	+	+	+	-				
BB	0	+								

virtue of additional *prior* information. We can thus identify each qualitatively specifiable elementary hypothesis by a unique number referable to a unique sampling distribution, viz. the total number of heads on the 4 faces of the two pennies chosen. Accordingly, we can lay out our hypotheses as an *ordered* set.

The Model IV (b) situation will here suffice to illustrate an essential step in the prescription of a proper recipe for stochastic induction, namely the discovery of an appropriate rule of exclusion and endorsement. The reader may find it helpful to explore the Model IV (a) situation in the same way. We shall assume that  $P_f < \frac{1}{14}$  is an acceptable uncertainty safeguard. Of itself, this is consistent with more than one rule of exclusion, as will be seen by comparison of Table I (in which the line corresponding to AB is irrelevant) with Table II in which the *positive* sign indicates a hypothesis retained as consistent with a particular set of observations and the *negative* sign indicates a hypothesis rejected as such because the risk of meeting such an observation if the hypothesis is true does not exceed  $\frac{1}{14}$ . In the phraseology of p. 392 the cells with the plus sign define the *endorsed* region of the stochastic H.O.G., the endorsed region of the non-stochastic grid including *also* the *excluded* cells labelled as such by the minus sign. For no hypothesis do the cells excluded by either rule specify more than a fraction equivalent to  $\frac{18}{256}$  (slightly less than  $\frac{1}{14}$ ) of our total experience consistent with its truth.

Table III, which summarises the outcome of two designs based on different rules of exclusion and endorsement, also shows what the outcome of the experiment may be in the non-stochastic domain. At a first glance we might yield to the temptation to say:

- (i) if our primary concern is to satisfy ourselves that neither penny is normal, we shall prefer the first rule of exclusion;
- (ii) if our primary concern is to satisfy ourselves that both pennies are normal, we shall choose the second.

If we did indeed base our preference on such considerations, we should not be interpreting the terms of reference of the uncertainty safeguard consistently. The latter applies to the rule in its

TABLE III

Observation No. of Heads	Non-Stochastic		Stochastic ( $P_f < \frac{1}{14}$ )		
	Type	(No. of Heads)	1st Rule of Exclusion		2nd Rule of Exclusion
			Type	No. of Heads	
0	CC, BC, BB	2, 1, 0	BB	0	BC, BB 0, 1
1	CC, BC	2, 1	BC	1	BC 1
2	CC, BC	2, 1	CC, BC	1, 2	CC, BC 1, 2
3	CC, BC	2, 1	CC, BC	1, 2	CC, BC 1, 2
4	BC, CC, AC	1, 2, 3	BC, CC, AC	1, 2, 3	CC 2
5	CC, AC	2, 3	CC, AC	2, 3	CC, AC 2, 3
6	CC, AC	2, 3	CC, AC	2	CC, AC 2, 3
7	CC, AC	2, 3	AC	3	AC 3
8	CC, AC, AA	2, 3, 4	AA	4	AC, AA 3, 4

*entirety* and has no bearing on the frequency of correct assertions limited to particular terminal statements. To discover a legitimate basis for preference we must therefore ask what the rule does accomplish in its entirety. Now Rule 2 ensures that no terminal statement endorses a composite hypothesis embodying more than 2 consecutive members of the comprehensive set of elementary hypotheses. Since the non-stochastic procedure is not *a priori adequate* to accomplish this, we may say that Rule 2 is conceivably acceptable in the *minimal* sense defined on p. 449; but Rule 1, like the non-stochastic procedure, does endorse a terminal statement referable to a composite hypothesis which encompasses 3 elementary hypotheses, viz. the assertion which goes with the observation of 4 heads in the 4-fold joint toss. Thus one lesson we learn from our model is that a prerequisite to prescription of a proper rule of stochastic induction is a Rule of Exclusion and Endorsement (R.E.E.) consistent *both* with the acceptable level of uncertainty which circumscribes its validity, *and* with the acceptable form of terminal statements it subsumes.

We shall also recognise an implication of the principle of *a priori adequacy* if we modify the foregoing procedure. Either rule of exclusion and endorsement we have explored is comprehensive in the sense defined on p. 346, i.e. we commit ourselves to a terminal statement referable to every observation we may make. If we are content to take the risk that as many as  $\frac{70}{256}$ —roughly a quarter—of our experiments will be fruitless in the long run, if 4 is the actual number of heads on the 4 faces of the 2 coins chosen in the Model IV (*b*) set-up and that *all* our experiments will then be fruitless in the Model IV (*a*) set-up, we are entitled to explore the possibility of prescribing a new design based on a greater number of joint tosses but still consistent with the condition  $P_f < \frac{1}{14}$ . What we cannot safely do is to adapt a comprehensive design to satisfy the requirements of *a priori adequacy* by disregarding a particular class of observations.

Since Rule 1 would guarantee the same overall precision of statement as Rule 2, in the Model IV (*b*) setting if we refrained from making any terminal statement referable to a head-score of 4 in the 4-fold joint toss, it is instructive to examine the

consequences of making our test non-comprehensive w.r.t. Rule I. Without changing the prescription in any other way, we shall suppose that we confine our attention to trials of which the outcome is any number of heads other than 4. If AA or BB correctly describe our choice the situation remains the same. If either AC or BC truly specifies our choice we shall confine our positive terminal statements to 15 out of 16 situations in our total experience. If the pair chosen is CC we decline to make a statement in 70 out of 256 situations and confine our positive terminal statement to 186 out of 256 in

TABLE IV

Hypothesis		Observation (No. of Heads)							
Type	No. of Heads	0	1	2	3	5	6	7	8
AA	4	0	0	0	0	0	0	0	1
AC	3	0	0	0	0	$\frac{4}{15}$	$\frac{6}{15}$	$\frac{4}{15}$	$\frac{1}{15}$
CC	2	$\frac{1}{186}$	$\frac{8}{186}$	$\frac{28}{186}$	$\frac{56}{186}$	$\frac{56}{186}$	$\frac{28}{186}$	$\frac{8}{186}$	$\frac{1}{186}$
BC	1	$\frac{1}{15}$	$\frac{4}{15}$	$\frac{6}{15}$	$\frac{4}{15}$	0	0	0	0
BB	0	1	0	0	0	0	0	0	0

our total experience. Accordingly, we must readjust the relevant figures of Table I, as in Table IV. We then see that 18 in 186 of our assertions will be false if the correct specification of the pair is CC. The overall operation of the rule now signifies that it is not inconsistent with  $\frac{1}{14}$  of total experience within the framework of the assumption that any single hypothesis may be true. Thus  $\frac{18}{186} > P_f$  as defined by the comprehensive rule. So the design will not be consistent with the acceptable level of uncertainty, if we reserve the right to restrict the terms of reference of the rule to observations consistent with the presumptive operational intent.



*Models V (a) and V (b).* So far, we have seen that different rules of exclusion and endorsement may be consistent with one and the same level of uncertainty, but only one of the two discussed above is also consistent with the criterion of *acceptable terminal statement*. That different laws of exclusion may fulfil both conditions will emerge from a study of two model situations with respect to each of which we shall postulate the following common features:

(a) a lottery wheel has  $N$  equal sectors each of which is black or red;

(b) the single-spin (unit trial) score depends on whether the sector which comes to rest opposite a fixed pointer is black (*zero* score) or red (*unit* score);

(c) we do not know the number ( $Np$ ) of red sectors or the number ( $Nq = N \cdot \overline{1-p}$ ) of black ones;

(d) we spin the wheel 400 times, and record the 400-fold *mean* score.

We now distinguish wheels of 2 sorts:

*Model V (a):* We know that:

- (i)  $N = 10$ ;
- (ii) at least one sector is red;
- (iii) at least one sector is black.

*Model V (b):* We know that:

- (i)  $N = 100$ ;
- (ii) at least 10 sectors are red;
- (iii) at least 10 sectors are black.

In either set-up we are sampling in a 2-class universe and successive terms of the binomial  $(q + p)^{400}$  define the sample mean score distribution. Of either set, we may also say that neither  $rq = 400q$  nor  $rp = 400p$  is less than 10, whence we may invoke (p. 160) the normal distribution as a satisfactory descriptive device. The essential difference between the two models is this. Of Model V (a) we know that  $p$  has one of a set of 9 *discrete* values from 0.1 to 0.9 in intervals of 0.1. Of Model V (b) we know that  $p$  has one of a set of 81 *discrete* values from 0.1 to 0.9 in intervals of 0.01.

Without assuming that we can actually design an experiment unless we assign in advance the sample size ( $r$ ) consistent with an acceptable set of terminal statements at an acceptable level of uncertainty, let us here examine the implications of arbitrarily fixing  $r = 400$  in terms of the adequacy of a design to endorse a statement about  $p$ . First we express in standard form the mean score values ( $p_{0.1}$  and  $p_{0.2}$ ) which define a range in which 95 per cent of all score values lie symmetrically about the mean  $p$ , viz.:

$$\frac{(p - p_{0.1})^2}{\sigma_p^2} = 4 = \frac{(p - p_{0.2})^2}{\sigma_p^2} \text{ and } \sigma_p^2 = \frac{p(1-p)}{400}$$

Thus  $p_{0.1} = 0.45$  and  $p_{0.2} = 0.55$  if  $p = 0.5$ . For any value of  $p$  other than 0.5 the interval consistent with the 5 per cent level of uncertainty will be shorter. Thus the probability that a mean score lies inside the range 0.35 to 0.45 when  $p = 0.4$  is less than 0.05.

In either set-up, the arbitrarily assumed value  $r = 400$  thus ensures that *more than* 95 per cent of all score values will lie inside the range specified below for each value of  $p$  cited:

range of mean score ( $p_0$ )	$p$
0.05 — 0.15	0.1
0.15 — 0.25	0.2
0.25 — 0.35	0.3
0.35 — 0.45	0.4
0.45 — 0.55	0.5
0.55 — 0.65	0.6
0.65 — 0.75	0.7
0.75 — 0.85	0.8
0.85 — 0.95	0.9

In the Model V (a) set-up these 9 values of  $p$  constitute a discrete comprehensive set of hypotheses and score values in the range defined by a rule of exclusion consistent with the uncertainty safeguard  $P_j \leq 0.05$  for each admissible hypothesis do not overlap. For Model V (a) we may thus make one rule of induction *en rapport* with decision test procedure embracing

a comprehensive set of 9 admissible hypotheses with an uncertainty safeguard  $P_f \leq 0.05$ , viz.:

If  $p_0$  is *inside* the range  $0.05 - 0.15$  we shall assert that  $p = 0.1$ .

If  $p_0$  is *inside* the range  $0.15 - 0.25$  we shall assert that  $p = 0.2$ .

If  $p_0$  is *inside* the range  $0.25 - 0.35$  we shall assert that  $p = 0.3$ .

*etc. etc. etc.*

For two reasons recognisable by reference to Fig. 2 the rule so stated is not comprehensive in the sense defined on p. 346:

(i) since the distribution of score values is discrete with fixed interval  $\Delta p_0 = 0.0025$  in the range 0 to 1, the restriction implied by *inside* in our statement of the rule excludes our right to make a statement when the value of  $p_0$  is an exact multiple of 0.05;

(ii) the rule permits us to make no statement if  $p_0 \leq 0.05$  or  $p_0 \geq 0.95$ .

Since the truth of any one of the 9 admissible hypotheses excludes the possibility that more than 5 per cent of samples we meet in the long run will be *null*, i.e. that we shall have to refrain from making statements about them, it happens that the disadvantages of (i) and (ii) in terms of economical design are trivial; but we can remove this minor disability in more than one way. First we recall that the interval which symmetrically circumscribes 95 per cent of all sample scores for all admissible values of  $p$  other than 0.5 is less than 0.1. Thus we shall not violate our uncertainty safeguard if we define our score interval for  $p = 0.5$  as 0.45–0.55 inclusive and each interval on either side symmetrically of length  $0.1 - 2\Delta p_0 = 0.995$ . Since the 95 per cent score range falls off from  $p = 0.5$  to  $p = 0.1$  and from  $p = 0.5$  to  $p = 0.9$ , we are free to accommodate the otherwise null score values in many *other* ways, each being an R.E.E. consistent with  $P_f \leq 0.05$ . Next we note that scores outside the range  $p_0 \leq 0.05$  or  $p_0 \geq 0.95$  must occur in less than 2.5 per cent of all samples even when  $p = 0.1$  or  $p = 0.9$ . So we shall not violate our uncertainty

safeguard if we define our terminal intervals with limits at  $p_0 = 0$  and  $p_0 = 1$ . Accordingly, we might state *one* rule consistent with an uncertainty safeguard  $P_f \leq 0.05$  and admitting no *null* observations as:

If $p_0$ lies <i>inclusively</i> in the range	Assert that $p$ is
0.00 — 0.1475	0.1
0.15 — 0.2475	0.2
0.25 — 0.3475	0.3
0.35 — 0.4475	0.4
0.45 — 0.55	0.5
0.5525 — 0.65	0.6
0.6525 — 0.75	0.7
0.7525 — 0.85	0.8
0.8525 — 1.0	0.9

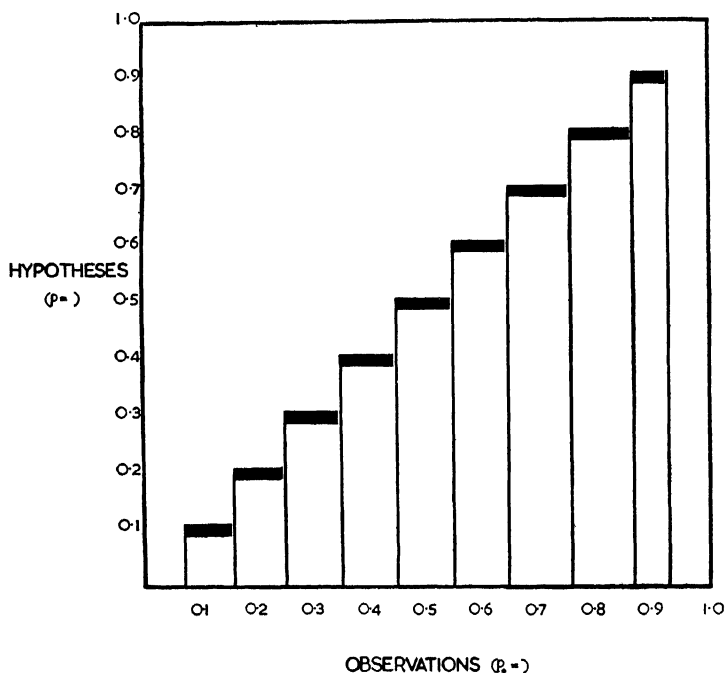


FIG. 2. Model V (a)

The length of each thick black horizontal line represents a 95 per cent range of all observations consistent with the corresponding hypothesis.

In the foregoing discussion of Model V (a) it has not been necessary to discuss what is an acceptable terminal statement. We have tacitly assumed that we wish to specify the exact value of  $p$  within the framework of an acceptable risk of error, and it transpires that the sample size  $r = 400$  satisfies the condition that our error risk is  $P_f \leq 0.05$ , i.e. does not exceed the conventionally prescribed level of uncertainty. If we demand the same precision of statement in the set-up of Model V (b), we shall need to prescribe  $r$  for a comparable test procedure, so that 95 per cent of score values lie symmetrically in an interval of  $0.01$  when  $p = 0.5$ , i.e. for a range bounded by  $p_0 = 0.495$  and  $p_0 = 0.505$ .

Whence we must define  $r$  so that

$$\frac{(0.5 - 0.505)^2}{\sigma_p^2} = 4 \quad \text{and} \quad \sigma_p^2 = \frac{1}{4r}$$

$$\therefore r = 40,000$$

An essential difference between the two models is therefore that a design to ensure correct acceptance of each admissible *elementary* hypothesis with an uncertainty safeguard  $P_f \leq 0.05$  prescribes recourse to samples 100 times as large for a Model V (b) as for a Model V (a) set-up. A sample size so large as 40,000 is too large to handle in almost any conceivable designed experiment. Accordingly, we must abandon hope if we identify each acceptable terminal statement as the assertion that a particular elementary hypothesis (discrete value of  $p$ ) is the true one; but we are then asking our design to guarantee a much higher level of numerical precision than the design for Model V (a) prescribed. In effect, we were then content to say that  $p$  lies in an *interval* of length  $0.1$ . Let us therefore now assume that assertions about  $p$  consistent with this precision level are acceptable terminal statements. Our set of *composite* hypotheses corresponding to 401 possible observations ( $p_0$ ) will then constitute a comprehensive set of 9 *consecutive overlapping intervals* in the sense defined above; and our procedure will offer no difficulty if we visualise these 401 possible observations and 81 elementary hypotheses laid out gridwise as a hypothesis-observation table with cell entries of a given row referable to

a particular elementary hypothesis and cell entries of a particular column referable to a particular observation (mean score), as for Type III of p. 428 below.

Each elementary hypothesis specified by one of the ordered set of parameter ( $p$ ) values in the range 0.1 to 0.9 with an interval  $\Delta p = 0.01$  admits the occurrence of an observation in each cell of the corresponding row labelled by mean score values ( $p_0$ ) of the 400-fold spin from 0 to 1 with an interval  $\Delta p_0 = 0.0025$ . Thus every cell of the grid would carry a positive sign, if we labelled it as for an experiment in the non-stochastic domain. No rule referable to single observations (mean score of 400-fold spin) in the domain of non-stochastic induction could then give us any information whatsoever about  $p$ . We may, however, take advantage of the fact that observations consistent with any elementary hypothesis are not of equal value in terms of the totality of our experience. Some are far more common than others, and we may black-out a class of rare ones according to a rule of exclusion which guarantees the same risk of erroneously stating that such and such a set of observations are inconsistent with the truth of such and such a hypothesis.

There will be many different ways of doing this, but the one which leads to the most tractable outcome is to exclude all values of  $p_0$  outside the symmetrical range  $p_i \pm h\sigma_{p_i}$  for every value of  $i$  which labels one or other of the elementary hypotheses. Our cell entries marked with the positive sign collectively then occupy a lozenge shaped (*endorsed*) region with a jagged outline running diagonally across the grid. If we interpret this as we should interpret a similar hypothesis-observation table for a non-stochastic domain, the column of marked cells corresponding to a particular score value will delimit the row marginal  $p$  values (elementary hypothesis) deemed to be consistent with the observation on the dual assumption that: (a) we operate the rule itself consistently; (b) we accept the assigned risk.

As before, we shall assign as an acceptable risk  $P_f \leq 0.05$ , so that  $h = 2$  above. To operate the rule consistently we must then ensure that it will guarantee the acceptable terminal statement, here expressible as a precision level. If we call the

marginal value of  $p$  referable to the uppermost cell of the  $j$ th column  $p_{v,j}$  and the one referable to the lowermost cell  $p_{u,j}$ , our precision level for the particular score value  $p_{0,j}$  will be  $(p_{v,j} - p_{u,j})$ . To define our rule in terms of an acceptable terminal statement we must specify an interval  $i_p$  so that  $(p_{v,j} - p_{u,j}) \leq i_p$ ; and this will be possible only if we define  $r$  appropriately for the  $r$ -fold spin. Thus we cannot arbitrarily

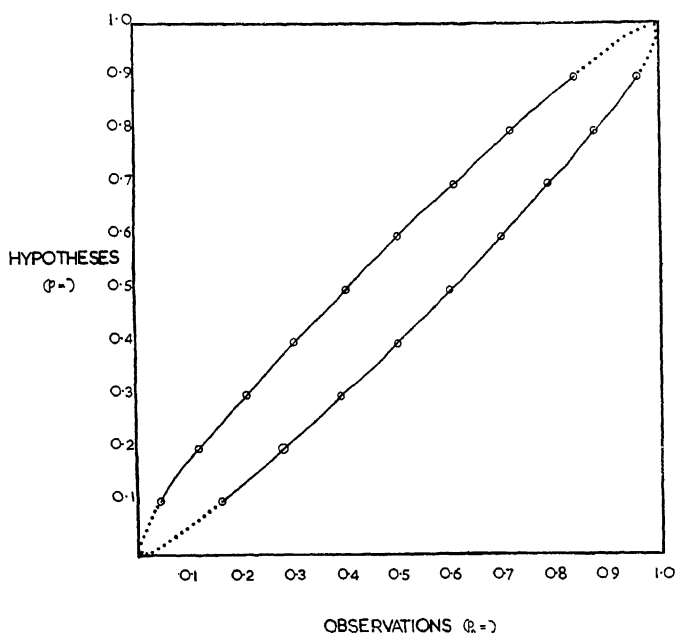


FIG. 3. Model V (b).

The Neyman solution.

assume that  $r = 400$  satisfies the criterion of acceptable terminal statement defined by  $i_p = 0.1$ . If we prescribe this as our criterion, we must seek to specify  $r$  appropriately.

When  $r$  is large—as we can satisfy ourselves by preliminary examination—we may provisionally regard the jagged outline of the *endorsed* region enclosing all admissible observations consistent with the prescribed rule of exclusion as two curves defined by the equations

$$p_{u,j} + 2\sigma_{p,v} = p_{0,j} = p_{v,j} - 2\sigma_{p,v} \quad . \quad . \quad . \quad (i)$$

As we know,  $\sigma_{p,j}$  has its maximum value when  $p = p_i = 0.5$  and the horizontal width of the endorsed region is greatest at this level. From Fig. 3 we see that the vertical width is also greatest for  $p_0 = 0.5$ . It will therefore suffice if we prescribe  $r$  to ensure that  $i_p \leq 0.1$  when  $p_0 = 0.5$  so that  $(p_{v,j} + p_{u,j}) = 1$ , whence  $p_{u,j} = (1 - p_{v,j})$  and  $(1 - p_{u,j}) = p_{v,j}$

$$\therefore \sigma_{p,v}^2 = \frac{p_{v,j}(1 - p_{v,j})}{r} = \frac{p_{u,j}(1 - p_{u,j})}{r} = \sigma_{p,u}^2$$

$$\therefore p_{v,j} - p_{u,j} = 4\sigma_{p,v} = 0.1$$

$$\therefore \sigma_{p,v}^2 = \frac{1}{1600}$$

Now  $\sigma_{p,v}^2$  is a maximum for  $p = 0.5$  when

$$\sigma_p^2 = \frac{1}{4r}$$

$$\therefore \frac{1}{1600} \leq \frac{1}{4r}$$

$$\therefore r \leq 400$$

This settles the question: is  $r = 400$  large enough to ensure an interval  $i_p \leq 0.1$  with an uncertainty safeguard  $0.05$ ? It does not locate the interval, nor does it tell us how small we may make  $r$ . Our uncertainty safeguard is  $P_f = 0.05$  for the particular case when  $h = 2$ . We may dispose of both issues last stated in more general terms than above, if we define  $P_f$  in terms of  $h\sigma$ . In that event we may write (i) above as

$$(p_{v,j} - p_{u,j})^2 = h^2\sigma_{p,v}^2 \text{ and } (p_{u,j} - p_{0,j})^2 = h^2\sigma_{p,u}^2$$

$$\therefore (r + h^2)p_{v,j}^2 - (2rp_{0,j} + h^2)p_{v,j} + rp_{0,j}^2 = 0$$

and

$$(r + h^2)p_{u,j}^2 - (2rp_{0,j} + h^2)p_{u,j} + rp_{0,j}^2 = 0$$

Formally, the last two equations are identical and their two roots are the roots of

$$(r + h^2)p^2 - (2rp_0 + h^2)p + rp_0^2 = 0$$



Thus we obtain

$$p_{v,j} = \frac{(2rp_{0,j} + h^2) + \sqrt{(2rp_{0,j} + h^2)^2 - 4r(r + h^2)p_{0,j}^2}}{2(r + h^2)} \quad (\text{ii})$$

$$p_{u,j} = \frac{(2rp_{0,j} + h^2) - \sqrt{(2rp_{0,j} + h^2)^2 - 4r(r + h^2)p_{0,j}^2}}{2(r + h^2)} \quad (\text{iii})$$

We wish to make  $(p_{v,j} - p_{u,j}) = 0.1$  when  $p_0 = 0.5$ , whence

$$\frac{1}{100} = \frac{4}{r + 4}$$

$$\therefore r = 396$$

For a given value of  $p_{0,j}$ , (ii) and (iii) give the appropriate limits of  $p$  consistent with the uncertainty safeguard. The solution of (ii) and (iii) for values of  $p_{0,j}$  other than 0.5 and a fixed value of  $r$  determined in this way to ensure the prescribed precision level, yields so-called confidence intervals  $(p_{v,j} - p_{u,j})$  of different length for each value of  $p_{0,j}$  on either side of  $p_{0,j} = 0.5$ . If  $p$  may have any values in the range from 0 to 1, we are therefore entitled to say that (ii) and (iii) subsume a rule consistent with an acceptable form of terminal statement and an uncertainty safeguard  $P_f = 0.05$  at the prescribed acceptable level. This is indeed the statement of the procedure advanced by Neyman's school; but it is open to three (see also p. 451) objections, two of which our Model IV brings sharply into focus.

First we note that  $rp$  and  $rq$  will both exceed 10, our criterion (p. 160) of an adequate normal fit, only if  $p$  lies in the range  $0.025 - 0.975$ , whence a method of assigning intervals of length less than the precision level (here 0.1) prescribed as the allowable maximum for  $p_{0,j} = 0.5$  will not be valid at the limit of the range of values  $p_{0,j}$  may assume. Next, let us ask what the consistent use of a rule circumscribed by (ii) and (iii) implies when  $p_{0,j}$  lies in the neighbourhood of zero or unity. For  $p_{0,j} = 0$ ,  $p_{v,j} = 0.01$  and  $p_{u,j} = 0$ , so that our terminal statement corresponding to the observation  $p_{0,j} = 0$  will be  $0 \leq p \leq 0.01$ . Similarly, we find  $0.99 \leq p \leq 1.0$  for  $p_{0,j} = 1$ . Both assertions are false in the Model V situation, since we

know that  $0.1 \leq p \leq 0.9$ , and in general consistent adherence to (ii) and (iii) must lead to false statements for some values of  $p_{0,j}$ , if background knowledge of prior possibilities limits the admissible range of the comprehensive set of hypotheses specified by values of  $p$ . That we shall always be wrong when we make terminal statements referable to observations outside the range stated is not inconsistent with what we can claim for the operation of the rule in its entirety; but it is not a circumstance which commends the rule to our good judgment.

*Model VI.* The two issues last stated do not explicitly force themselves on our attention when we assume that the universe of choice is normal. We shall defer their consideration to Chapter 18 in which we shall examine the need for a modification of the theory of interval estimation prescribed by Neyman. First, we may clarify some implications of a rule of stochastic induction when we assume a randomwise sampling process in a putative normal universe by exploring the following set-up.

A lottery wheel has 1024 sectors labelled with scores  $x$ ,  $(x + 1)$ ,  $(x + 2)$ ,  $(x + 3)$  . . .  $(x + 9)$ ,  $(x + 10)$  respectively allocated to 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1 sectors. We do not know the numerical value of  $x$ , but know that it is any one of a set of consecutive positive numbers in the range  $0 - 10$  with fixed interval  $\Delta x = 0.01$ . The long-run mean value ( $M$ ) of the score of any sample is, of course  $(x + 5)$ ; and the terms of  $(\frac{1}{2} + \frac{1}{2})^{10}$  define the unit sample distribution of the universe with variance  $\sigma^2 = 2.5$ , whence that of the distribution of the 40-fold sample mean is

$$\sigma_m^2 = \frac{\sigma^2}{40} = \frac{1}{16}$$

Thus  $\sigma_m = 0.25$ ; and the error involved in a normal quadrature for the distribution of the sample means is trivial. We can thus say that

- (a) the mean ( $M_x$ ) of 2.5 per cent of all samples in the long run will exceed  $M + 2\sigma_m = M + 0.5$ ;
- (b) the mean of 2.5 per cent of all samples in the long run will be less than  $M - 2\sigma_m = M - 0.5$ ;
- (c) the mean of 95 per cent of all samples will lie in the range  $M \pm 2\sigma_m = M \pm 0.5$ .

# RECIPE AND RULE IN STOCHASTIC INDUCTION

In this set-up the slightly jagged outlines of the region of the H.O.G. endorsed by the rule closely follow two parallel straight lines, whence the length of the interval (vertical width of the endorsed region) referable to each admissible observation will be the same for each, as illustrated in Fig. 4. There we see that the slope of the two bounding lines is unity, whence the vertical and horizontal widths of the endorsed region are identical. In this set-up an R.E.E. consistent with  $P_f \leq 0.05$  is also consistent

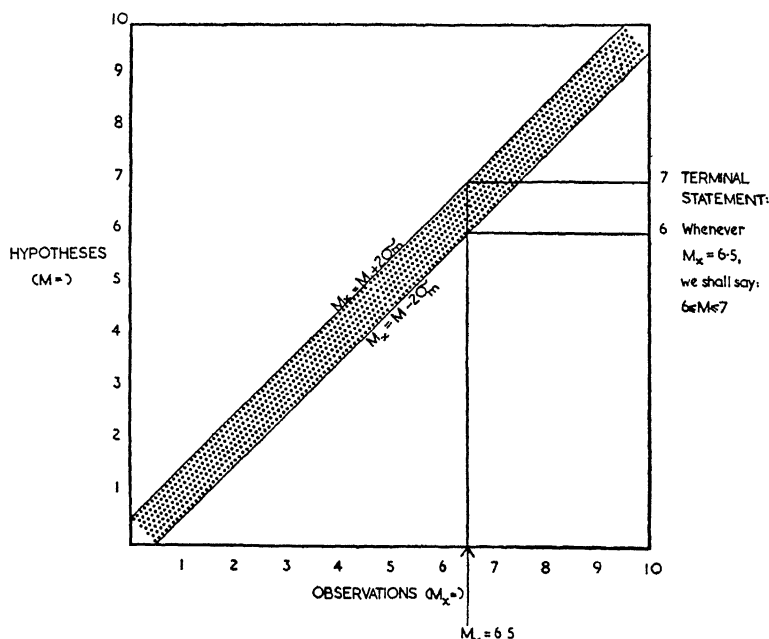


FIG. 4. Model VI.

with the assumption that our criterion of acceptability w.r.t. the set of terminal statements is the precision level  $i_m = 1.0$ . If the sampling distributions were truly normal, we could set no limits on the range of our observations, and we could disregard extreme values of  $M_x$  with impunity. It would then be possible to write  $P_f = 0.05$  in this context.

Now the choice of our criterion of acceptability ( $i_m = 1.0$ ) is quite arbitrary. If we are to acquaint ourselves *with all the steps we must traverse* when we seek to prescribe a proper rule of

stochastic induction, we must therefore ask how we should define the numerical value of  $r$  appropriate to any agreeable criterion of acceptability w.r.t. both the risk and the precision level. We then write

$$i_m = 2h\sigma_m \text{ and } \sigma_m^2 = \frac{\sigma^2}{r} = \frac{2.5}{r}$$

$$\therefore r = \frac{10h^2}{i_m^2}$$

If, as above,  $P_f \leq 0.05$ , so that  $h = 2$ , our complete specification of a rule of stochastic induction to guarantee precision levels cited will be to assign as the sample size (i.e. number of times we spin the wheel) the values of  $r$  shown

$i_m$	$r$
1.0	40
0.5	160
0.25	640
0.1	4,000

The calculations involved in this essential step are possible because we know the true value  $\sigma^2$  of the variance of the u.s.d. referable to the sampling distribution definitive of each elementary hypothesis specified by an admissible value  $M$  may have. We get to the root of one of the difficulties which has embarrassed an adequate statement of the theory of interval estimation fully consistent with the Forward Look if we now ask what we might say, if we did *not* possess such knowledge. In such circumstances each sample we take gives us an estimate  $s_m$  of  $\sigma_m$ . From the  $t$ -distribution we then have:

$$(i) \quad \frac{M_x - M}{s_m} = t ; \quad (ii) \quad s_m^2 = \frac{1}{r(r-1)} \sum_{u=1}^{u=r} (x_u - M_x)^2$$

$$(iii) \quad P(M - h.s_m \leq M_x \leq M + h.s_m) = \int_{-h}^h F(t) dt$$

When  $r = 12$  the  $t$ -table, i.e. the table of the integral of  $F(t)$ , gives  $h = 2.2$  for  $P(M - h.s_m \leq M_x \leq M + h.s_m) = 0.95$ .

It is therefore tempting to say that  $P_f = 0.05$  is the uncertainty safeguard for the operation of the rule:

say that  $M_x + 2.2s_m \geq M \geq M_x - 2.2s_m$  whenever we take a 12-fold sample for whatever values of  $M_x$  we may encounter.

If so, our calculation will proceed as follows for a 12-fold sample of score values:

2, 7, 9, 9, 12, 18, 21, 33, 37, 44, 51, 80

These yield  $M_x = 26.92$  and  $s_m \simeq 6.64$ . Seemingly, our so-called rule will therefore lead us to assert:  $12.31 \leq M \leq 41.53$  with uncertainty safeguard 0.05 when  $M_x = 26.92$ . Let us now suppose we take a second sample constituted thus:

1, 3, 8, 8, 8, 10, 20, 25, 29, 32, 49, 130

Again  $M_x = 26.92$  but  $s_m \simeq 10.22$  and our so-called rule leads us to assert  $4.44 \leq M \leq 49.40$ .

At first sight, we may be inclined to dismiss this discrepancy since: (a) the two values of  $M_x$  are referable to samples of different make-up; (b) the use of the statistic  $s_m$  in the definition incorporates the relevant difference. What is not at once apparent is that our reliance on the latter excludes our right to prescribe a one-stage trial for which we can preassign the sample size  $r$  consistent *both* with the acceptable level of uncertainty *and* with a criterion of acceptable terminal statement. Though  $M_x$  and  $s_m$  incorporate all the information a sample can contribute to our knowledge of the parent universe, they do not incorporate singly or jointly the information we require if we are to complete the formulation of a rule of stochastic induction in its entirety.

In the ideally normal domain of the pure mathematician an infinitude of values of  $\sigma$  are consistent with a single value of  $M$ , and for each of these the single  $r$ -fold distribution of  $(M_x - M)$  in the row of infinitesimal elements our hypothesis-observation grid can accommodate will be different. In this ideally normal domain,  $s_m$  and  $M_x$  are independent variables and each admits an infinitude of values. Whence the range of the  $t$ -ratio for each pair of admissible values of  $M$  and  $\sigma$  for a fixed value of  $r$

is also infinite. In short, the mathematical theory of the  $t$ -distribution, albeit admittedly and at best a rough and ready approximation to experience of the real world, endorses no guarantee that our confidence interval, as defined by Neyman (p. 437) or the fiducial interval as defined by R. A. Fisher below (p. 441), will necessarily be finite. If not, our so-called rule violates the criterion of minimal acceptability of terminal statement stated on p. 449. That is to say, it cannot ensure that the outcome of applying it will endorse a more precise statement than we might justify by reasoning in the domain of non-stochastic induction.

Against the background of the lottery wheel of the Model VI set-up, we may clarify the inadequacy of the  $t$ -distribution to provide a basis for interval estimation in a one-stage trial in the following terms. To prescribe a rule of stochastic induction involving a comprehensive set of acceptable terminal statements referable to composite hypotheses, two steps are necessary:

(i) to define a rule of exclusion and endorsement which will guarantee either of the two following:

(a) an acceptable uncertainty safeguard for a fixed sample size when we have not as yet fixed our criterion of acceptable terminal statement;

(b) an acceptable precision level when we have not as yet fixed our risk of erroneous statement;

(ii) to define the size ( $r$ ) of sample requisite to ensure the acceptable precision level when we have formally stated an R.E.E. consistent with a criterion of acceptability w.r.t. the set of terminal statements.

There is as yet no general principle which subsumes (i) and (ii) for all appropriate situations. Nor is it likely that we shall be able to accomplish the end in view without recourse to hit-and-miss procedures, when we come to grips with a new one. Be that as it may, the inadequacy of the  $t$ -distribution resides in the circumstance that it provides an exploratory

basis only for (a) stated above as only one of two essential steps in the prescription of a rule of stochastic induction. It cannot help us to proceed by the alternative route and it cannot help us to take the second equally essential step.

*Patterns of Stochastic Induction.* The device of the H.O.G. sheds a new light on a familiar type of graph such as Fig. 3 illustrating Neyman's theory of interval estimation for a fixed sample size. We can interpret it rightly as a visualisation of the *first* step towards the prescription of a rule of procedure which is complete only when we take the next step by prescribing the size of sample consistent with an acceptable criterion of terminal statement. In the enumerative domain the number of cells is necessarily finite; but we can make the transition from grid to graph with little effort, when we pass into the domain of continuous sampling distributions. Accordingly, it may be instructive if we pause at this stage to examine the guises the H.O.G. may assume both in the domain of non-stochastic and of stochastic induction.

PATTERNS OF STOCHASTIC INDUCTION. We have seen that we can specify the elementary hypotheses of the Model IV (b) set-up in purely qualitative terms or in numerical terms consistent with the lay-out of an ordered set. When we can do so, we may helpfully distinguish between two situations in a set-up which does admit of a precise qualitative definition of hypotheses: (a) the ordered set is arbitrary in the sense that the qualitative specification of the hypotheses admits of no intelligible and/or useful corresponding arrangement; (b) the ordered set at least corresponds to an intelligible and/or useful system of *rank* scoring, e.g. black, brown, yellow, cream, white. If (b) holds good, the problem of experimental design is on all fours with what arises in situations admitting no qualitative specification of elementary hypotheses, as is true of the urn model introduced on p. 390 to illustrate how stochastic differs from non-stochastic induction. That is to say, we can define composite hypotheses with each of which we can associate a terminal statement referable to a range or interval.

Whenever we can make an intelligible and/or useful ordered sub-classification of elementary hypotheses, we can draw an

intelligibly sharp distinction between two ways in which we can do so: (i) no members of a sub-set corresponding to a particular acceptable terminal statement occur in any other such subset; (ii) each acceptable terminal statement refers to a subset of which some members occur in others. We may then speak of the composite hypothesis corresponding to each terminal statement as a *consecutive discrete interval*. If (ii) is admissible, it may be possible to grade the subsets so that successive subsets contain an ordered terminal sequence in common with an ordered terminal sequence of a successor and predecessor. We may then speak of the composite hypothesis corresponding to each terminal statement as a *consecutive overlapping interval*. Among possible patterns of the hypothesis-observation table of an experimental design distinguishable in terms of (i) and (ii) above for 8 elementary hypotheses and 4 observations (score values) are those shown in Fig. 5. Only the upper two labelled as (i) are consistent with *a priori* adequacy of design in the non-stochastic domain.

The two types of arrangement shown opposite do not exhaust all ways in which we might conceivably combine elementary hypotheses into composite hypotheses corresponding to acceptable terminal statements. The usefulness of grouping elementary hypotheses in one way or another depends on the acceptable terminal statements consistent with the design; but a stochastic design gives us a greater freedom to combine elementary hypotheses into subsets consistent with the whole operational intent. The reason for this is that the rule of exclusion consistent with the acceptable level of uncertainty permits us to disregard some elementary hypotheses consistent with a particular observation. Thus it is often possible to design an experiment which is *a priori adequate* in the stochastic domain for a *particular set* of acceptable terminal statements when it would not be possible to do so otherwise. In the domains of Models V (a) and (b) and of VI, every possible observation is—if only very rarely—realisable within the framework of the presumptive truth of every admissible elementary hypothesis. Thus no rule of non-stochastic induction is admissible since no admissible terminal statements could ever be acceptable.

If we now look at possible patterns more explicitly in terms



# RECIPE AND RULE IN STOCHASTIC INDUCTION

(i)

	0	1	2	3
1	+			
2	+			
3		+		
4		+		
5			+	
6			+	
7				+
8				+

	0	1	2	3
1	+			
2		+		
3		+		
4		+		
5			+	
6			+	
7				+
8				+

(ii)

	0	1	2	3
1	+			
2	+	+		
3	+	+	+	
4		+	+	+
5			+	+
6				+
7				+
8				+

	0	1	2	3
1	+			
2	+	+		
3		+	+	
4			+	
5			+	+
6				+
7				+
8				+

FIG. 5

of stochastic induction, when the elementary hypotheses are numerically and meaningfully specifiable as an ordered set, we may classify types of H.O.G. mentioned in this and the last chapter as follows.

*Type I.* Each acceptable terminal statement endorses a discrete parameter value.

(a)

$x =$	0	1	2	3	4	5
$p = p_a$	+	+	+	+	-	-
$p = p_b$	-	-	-	-	+	+

*Rule of Exclusion one-sided*

*Terminal statements:*

- (i) if  $x \leq 3$  say  $p = p_a$ ,      (ii) if  $x > 3$  say  $p = p_b$ .

Cf. *Drosophila* Model, p. 348.

# RECIPE AND RULE IN STOCHASTIC INDUCTION

(b)

$x =$	0	1	2	3	4	5	6	7	8	9
$p = p_1$	—	+	+	+	—	—	—	—	—	—
$p = p_2$	—	—	—	—	+	+	+	+	+	—
$p = p_3$	—	—	—	—	—	—	—	—	—	+
$p = p_4$	—	—	—	—	—	—	—	—	—	—

$x =$	10	11	12	13	14	15	16	17	18
$p = p_1$	—	—	—	—	—	—	—	—	—
$p = p_2$	—	—	—	—	—	—	—	—	—
$p = p_3$	+	+	+	—	—	—	—	—	—
$p = p_4$	—	—	—	+	+	+	+	—	—

*Rule of Exclusion two-sided*

*Terminal statements:*

- (i) if  $x \leq 3$  say  $p = p_1$ ,      (iii) if  $8 < x \leq 12$  say  $p = p_3$ ,  
(ii) if  $3 < x \leq 8$  say  $p = p_2$ ,      (iv) if  $x \geq 13$  say  $p = p_4$ .

Cf. Model V (a) above.

*Type II.* The acceptable terminal statements endorse alternative composite hypotheses and each acceptable terminal statement is consistent with a range of different observational values.

(a)

$x =$	0	1	2	3	4	5	6
$p_1$	+	+	+	+	-	-	-
$p_2$	+	+	+	+	-	-	-
$p_3$	+	+	+	+	-	-	-
$p_4$	+	+	+	+	-	-	-
$p_5$	+	+	+	+	-	-	-
$p_6$	-	-	-	-	+	+	+
$p_7$	-	-	-	-	+	+	+
$p_8$	-	-	-	-	+	+	+

*Rule of Exclusion one-sided*

*Hypothesis  $H_1$ :*  $p$  has one of the values

$p_1, p_2, p_3, p_4, p_5$

*Hypothesis  $H_2$ :*  $p$  has one of the values

$p_6, p_7, p_8$

*Terminal statements:*

- (i) If  $x \leq 3$  assert  $H_1$  is true,
- (ii) If  $x > 3$  assert  $H_2$  is true.

(Cf. the dilemma of test procedure in the domain of the clinical trial, p. 367.)

(b)

$x = \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$

$p_1$	+	+	+	+	-	-	-	-
$p_2$	+	+	+	+	-	-	-	-
$p_3$	+	+	+	+	-	-	-	-
$p_4$	+	+	+	+	-	-	-	-
$p_5$	+	+	+	+	+	+	+	+
$p_6$	+	+	+	+	+	+	+	+
$p_7$	+	+	+	+	+	+	+	+
$p_8$	-	-	-	-	+	+	+	+
$p_9$	-	-	-	-	+	+	+	+
$p_{10}$	-	-	-	-	+	+	+	+

*Rule of Exclusion one-sided*

*Hypothesis  $H_1$ :  $p$  has one of the values*

*$p_1, p_2, p_3, p_4, p_5, p_6, p_7$*

*Hypothesis  $H_2$ :  $p$  has one of the values*

*$p_8, p_9, p_{10}$*

*Terminal statements:*

- (i) If  $x \leq 3$  assert  $H_1$  is true,
- (ii) If  $x > 3$  assert  $H_2$  is true.

(Cf. Wald's consumer-producer risk procedure, pp. 367-8.)

(c)

$x =$	0	1	2	3	4	5	6
$p_1$	+	+	-	-			
$p_2$	-	-	+	+	+	-	-
$p_3$				-	-	+	+

*Rule of Exclusion two-sided*

*Terminal statements:*

- (i) if  $x \leq 1$  say  $p = p_1$ ,
- (ii) if  $2 \leq x \leq 4$  say  $p = p_2$ ,
- (iii) if  $x > 4$  say  $p = p_3$ .

(Cf. Model IV (b) above.)

*Type III.* The acceptable terminal statements are composite hypotheses and each observation is referable to a different set of elementary hypotheses (see opposite page).

Types I to III do not exhaust all possible patterns of the stochastic hypothesis-observation grid conceived in terms of discrete score values (observations) and discrete values of  $p$  (hypotheses). As a condition of a manageable size of sample to endorse a commonly acceptable level of uncertainty, Type I presupposes that the interval between any two successive values of  $p$  is small. If  $p$  has only 2 admissible values, the appropriate rule of exclusion is necessarily one-sided. If  $p$  has more than 2 admissible values, a two-sided rule of exclusion will commonly be most easy to explore as for the urn model of p. 151 and for Model III in this context; but Model II in this context invokes both methods of exclusion. Type II (a) embraces Type I (a) as a special case. Its interest arises from the fact that it is unworkable when  $p$  admits a continuous range of values, since no acceptable sample size ( $r$ ) is then consistent with an acceptable level of uncertainty. Type II (a)

# RECIPE AND RULE IN STOCHASTIC INDUCTION

x = 0 1 2 3 4 5 6 7 8 9 10 11 12

P <sub>1</sub>							-	-	-	+	+	-
P <sub>2</sub>							-	-	-	+	+	-
P <sub>3</sub>						-	-	-	+	+	+	-
P <sub>4</sub>				-	-	+	+	+	+	+	+	-
P <sub>5</sub>			-	-	+	+	+	+	+	+	-	-
P <sub>6</sub>			-	-	+	+	+	+	+	+	-	-
P <sub>7</sub>		-	-	+	+	+	+	+	+	+	-	-
P <sub>8</sub>		-	-	+	+	+	+	+	+	-	-	-
P <sub>9</sub>	-	-	+	+	+	+	+	+	-	-	-	-
P <sub>10</sub>	-	-	+	+	+	+	+	-	-	-	-	-
P <sub>11</sub>	-	-	+	+	+	+	-	-	-	-	-	-
P <sub>12</sub>	-	-	+	+	+	+	-	-	-			
P <sub>13</sub>	-	+	+	+	+	-	-	-	-			
P <sub>14</sub>	-	+	+	-	-	-	-	-	-			

## Rule of Exclusion two-sided

Comprehensive (minimal) terminal statement:

If the interval  $p_{h+1} - p_h = \Delta p$ , and  $p_x$  is the median value of  $p_h$  corresponding to a particular observation ( $x_h$ ):

if  $x - x_h$  say  $p_x - 4\Delta p < p_h < p_x + 4\Delta p$

. . . . .

which is Wald's overlapping dual test procedure (p. 368) side-steps this difficulty by defining the upper limit of  $p$  for one composite hypothesis and the lower limit of  $p$  for the alternative at the price of excluding the possibility of making a statement asserting that  $p$  lies between these limits. Type III is the pattern of what we commonly mean by interval estimation. The rule of exclusion is necessarily 2-sided. If we compare Types I ( $b$ ) referable to such a model as V ( $a$ ) above with Type III referable to Models V ( $b$ ) or VI or Type I ( $a$ ) with Type II ( $b$ ), we are entitled to regard the distinction between test procedure and interval estimation as less important than the distinction between procedures which involve one-sided, two-sided or mixed rules of exclusion and endorsement.

*Rules of Induction.* We shall more readily face our next task (Chapter 17), if we now recapitulate the outcome of our examination of induction, whether stochastic or non-stochastic, in this chapter and in the last. Induction of either sort subsumes rules which endorse a decisive outcome whenever applied, if and only if: ( $a$ ) the hypotheses are comprehensive in the sense that they embrace all admissible contingencies in the domain of observation; ( $b$ ) the blueprint is *a priori adequate* in the sense that each observation endorses only one of a unique comprehensive set of acceptable terminal statements each of which itself endorses only one of a unique comprehensive set of hypotheses. In the non-stochastic domain, the specification of hypotheses which constitute a comprehensive set for design of a holonomic experiment may be expressible only in qualitative terms. In the stochastic domain they may be definable in such terms, but they *must* always be also expressible, if distinguishable, in terms of unique parameter values relevant to the specification of an appropriate sample distribution. What we may prefer to speak of as an elementary hypothesis or as a composite hypothesis in either domain is a matter of no importance except in so far as the distinction is relevant to: ( $a$ ) specification of the class of terminal statements deemed to justify the design of the experiment; ( $b$ ) the uniqueness of the relevant sample distribution.

In non-stochastic induction we formulate a rule of procedure which should: ( $a$ ) never lead to a false conclusion; ( $b$ ) always



lead to a definite and acceptable conclusion, if the design is *a priori* adequate. In stochastic induction we formulate a rule of procedure with which we can associate a numerically specifiable upper limit to the risk of erroneous statement if the design is *a priori* adequate; but the design must then encompass a specification of the number of requisite observations, if we wish to fix the risk at a level agreed as acceptable. This risk is not referable to *any individual* terminal statement or class of terminal statement. It is referable only to the consequences of applying the rule *in its entirety*. To speak of a rule which we must operate in its *entirety* presupposes:

- (a) that the set of hypotheses we are exploring is in fact comprehensive;
- (b) that the design is *a priori adequate* for the set of terminal statements the rule endorses.

It goes without saying that we accomplish nothing by accepting a risk of error, if we can thereby say no more than our observations would otherwise entitle us to say. Consequently, we may enlarge our previous comparison in the following terms:

(i) In the domain of non-stochastic induction, we can explore situations which suggest new hypotheses to test *pari passu* a comprehensive set of hypotheses. In stochastic induction we must state the set *in advance*. Whence non-stochastic induction embraces experiments which are exploratory as well as experiments which are holonomic. Contrariwise, stochastic induction is admissible only in the holonomic domain.

(ii) In the domain of non-stochastic induction, we may be content if the enquiry leads to any addition to our knowledge, whether by design or luck. In the domain of stochastic induction, our criterion of an acceptable terminal statement must at least guarantee at the outset a verdict which is more informative than whatever verdict we might reach by relying on non-stochastic induction.

Having given reasons which seem to me to be sufficiently compelling, I shall now put forward dogmatically the following canons illustrated by our discussion of model situations in this chapter and in the last.

*Canon 1*

A necessary condition for prescribing a rule of stochastic induction is that we can specify for each elementary hypothesis endorsed as admissible by one of the set of acceptable terminal statements a *unique* sampling distribution of admissible observations.

*Canon 2*

The legitimate use of a rule of stochastic induction in terms of an assigned and acceptable risk of error presupposes its operation in its entirety; and excludes the right to make any statement of the form: if we meet the *particular* observation  $x$  the probability that the particular elementary or composite hypothesis  $H_p$  is true is  $P_t$ . Instead we say:  $P_f = (1 - P_t)$  is the risk of erroneous statement if we *always* adhere to the rule regardless of whatever value  $x$  may have on a particular occasion.

*Canon 3*

That more than one rule of stochastic induction with one and the same uncertainty safeguard may be consistent with the same comprehensive set of elementary hypotheses and with the same set of acceptable statements is fully consistent with the acceptance of the preceding canons, and is a necessary consequence of our freedom to prescribe different rules of exclusion at the same level of risk.

*Canon 4*

The complete statement of a rule of stochastic induction must embrace:

- (a) a criterion of *acceptability* for the entire class of terminal statements subsumed;
- (b) an *acceptable risk* of erroneous statement;

When sampling occurs in one stage, (a) and (b) jointly signify that the rule must embrace a specification of the *size of the sample*.

## CHAPTER EIGHTEEN

### THE CONFIDENCE CONTROVERSY AND THE FIFTH CANON

IN THE LAST TWO CHAPTERS I have discussed at length a novel viewpoint of Wrighton, because it will help us to retrace our steps from a quagmire of controversy over interval estimation to a solid foothold for a new approach. That writers on statistical theory continued to speak of Fiducial limits and Confidence limits interchangeably throughout a decade after the relevant original publications first appeared should not surprise us, if we realise the difficulty of the task of extricating a novel restatement of principle from an overgrowth of custom thought embodied in current practice. With respect to the formal issue which divides the followers of Neyman from those of Fisher, the approach we have explored in the last two chapters is unequivocally opposite to the fiducial argument, and starts from the same premises as the theory of Confidence intervals; but it leads us to conclusions which do not emerge explicitly in Neyman's own writings, and the outcome will be to accommodate some objections which his critics have legitimately advanced.

It will help us to do justice to the disputants in the current controversy before we proceed to examine more closely Wrighton's own recent restatement, if I cite at length an exposition of his own position by Neyman himself.

In *Foundations of the General Theory of Statistical Estimation*,\* Neyman says:

. . . a brief summary of the general ideas underlying the theory of statistical estimation which I developed about 1930 and which, more recently, were brilliantly extended and generalized by Abraham Wald and by many other authors . . . it may be useful to mention briefly in what ways the new theory of estimation, or, more precisely, the theory of estimation which was new in the 1930's, differs from the earlier methods of attacking the problem. This difference may be symbolized by the change in labels.

\* xviii *Congrès Intern. Phil. Sci.* (1949).

Previously, attempts to build up a theory of statistical estimation went under the label "inductive reasoning." On the other hand, my own attempt was characterized by the label "theory of inductive behavior." The difference behind these two labels is primarily a difference in attitude towards the problem of estimation. . . . The approach to the theory of estimation characterized by the label of inductive reasoning consists in the search for such values ascribable to the parameters to be estimated, which, for one reason or another, are supposed to inspire the greatest possible confidence. The typical question asked would be :

in these circumstances, when these particular values of the random variables under consideration have been observed, what are the most probable (or the most likely) values of the unknown parameters?

Contrary to this, the more recent approach to the problem of estimation does not deal with the relative probability or likelihood of the various values ascribable to the unknown parameters but with the frequency of errors which will be committed if this or that particular method of estimation is consistently applied. In order to formulate the typical questions of this theory we introduce some notation. Let one letter  $X$  stand for the whole system of observable random variables with their joint distribution depending upon an unknown parameter  $\theta$  and let  $f(X) \leq g(X)$  be some two functions of  $X$ . The original approach to the problem of estimation from the point of view of inductive behaviour contemplates a process of estimation which consists (a) in observing the values, say  $x$ , of the random variables  $X$  and (b) in asserting that the true value of the parameter satisfies the double inequality  $f(x) \leq \theta \leq g(x)$ . The typical questions of the new theory are asked *before* the values  $x$  of the random variables  $X$  are observed. The first question is :

(i) Given the distribution of  $X$  and given the two functions  $f(X)$  and  $g(X)$ , what is the probability that the assertion  $f(X) \leq \theta \leq g(X)$  will be correct?

The second question is more important and much more interesting :

(ii) Given the distribution of  $X$  and given a number  $\theta$  between zero and unity, can one determine the two functions  $f(X)$  and  $g(X)$  so that the probability that the assertion  $f(X) \leq \theta \leq g(X)$  is true equals  $\alpha$  (or is at least equal to  $\alpha$ ), irrespective of the unknown value of the parameters and also irrespective of other parameters on which the distribution of  $X$  may depend?

. . . If the two functions  $f(X)$  and  $g(X)$  satisfy the conditions described in question (ii) then they are called *confidence limits* for estimating  $\theta$ , corresponding to the confidence coefficient  $\alpha$ . Also the random interval  $[f(X), g(X)]$  is called the *confidence interval*. . . . The reason for abandoning the old point of view on the problem of estimation should have been apparent for about the last century and a half, since the time when the formula of Bayes was discovered. The reason is that in order to obtain the *a posteriori* most probable values of the unknown parameters one must use the formula of Bayes which depends explicitly on probabilities *a priori* which are not implied by the circumstances of many problems where statistical estimation is needed. . . . To remedy the situation, some authors tried, and still try, to conjure the missing probabilities *a priori* out of a specially devised postulate, which I prefer to call a dogma. Some other authors dislike this particular postulate and give up the idea of "most probable" values of the parameters. Instead, they produce a recipe for constructing a function to measure our confidence in any stated value ascribable to a parameter, given the results of the observations, and then look for "optimum" values of the parameter.

Such methods of attack appear to me as evasions rather than solutions of the problem of estimation. Since about 1930, it was clear to me that, in order to attain the solution, only two methods were open: either find an error in the formula of Bayes (and I most sincerely believe that this is impossible) or modify the mathematical problem behind the methods of estimation so as to make it soluble in more general cases.

The reader will recognise that the formulation of Neyman's concept of inductive behaviour is fully *en rapport* with the viewpoint referred to as the *Forward Look* throughout this book; and registers a reorientation towards the proper terms of reference of a calculus of judgments explicitly stated in terms consistent with a behaviourist outlook. To Neyman first and foremost is due the credit for this reorientation; but it is singularly unfortunate that his views took shape at a time when the engaging algebraic properties of the *t*-distribution (p. 419) occupied the centre of the arena of statistical discussion, and seemed to offer a prospect of sidestepping uncertainties which had previously dogged the discussion of estimation and test procedure. Had he initially explored the implications of a behaviourist viewpoint in the discrete domain of classical models, as in his recently

published volume elsewhere cited, he would have seen what we have seen in Chapter 17. To speak of an act of will involved in following a rule of stochastic induction conveys little unless we assume a target of our endeavours. It is then necessary both to formulate a procedure with which we can associate an acceptable risk and one which guarantees an acceptable set of terminal statements. Neyman's formulation cited in this context violates our Canon 4 (p. 432), because it defines the rule to follow in terms of the acceptable risk alone; and this leads to an impasse in the field of the first application of his theory of interval estimation, viz. a hypothesis-observation set-up in which:

(a) the comprehensive set of hypotheses is a continuum of values the true mean ( $M$ ) of a normal universe may conceivably have;

(b) the observations constitute a continuum of sample mean values ( $M_x$ );

(c) the terminal statements vindicate composite hypotheses, each embodying a particular range of values of  $M_x$  within which  $M$  lies.

We have indeed (pp. 419-420) already explored the implications of the Forward Look *vis-à-vis* the usefulness or otherwise of the  $t$ -distribution to accomplish (c) on the assumption that the parent universe is normal; but it will not be profitless if we now re-examine it as case material to illustrate the *casus belli* of the confidence controversy. In the symbolism of p. 418, we may express in the most general form as below the probability that the sample mean ( $M_x$ ) lies in an interval uniquely determined by the unknown true mean ( $M$ ), the unbiased sample estimate ( $s_m$ ) of the variance of the mean and two arbitrary constants ( $h$  and  $k$ ) which we are free to fix at will:

$$P(M + k.s_m \leq M_x \leq M + h.s_m) = \int_k^h F(t).dt \quad (i)$$

In this equation the expression on the right is the tabulated integral of the  $t$ -distribution. If we wish to make our interval

symmetrical about the mean, we shall then write  $k = -h$ , so that

$$P(M - h \cdot s_m \leq M_x \leq M + h \cdot s_m) = \int_{-h}^h F(t) \cdot dt = P_t \quad (\text{ii})$$

Alternatively, we may wish to specify the probability that  $M_x$  will not exceed the limit determined by  $h$  for samples specified by the independent statistic  $s_m$ , i.e.

$$P(M_x \leq M + h \cdot s_m) = \int_{-\infty}^h F(t) \cdot dt \quad . \quad . \quad (\text{iii})$$

The foregoing statements involve no departure from the classical approach. In Neyman's statement of the problem of interval estimation, we proceed therefrom as follows:

(a) when  $(M - s_m \cdot h \leq M_x \leq M + s_m \cdot h)$  it will be true that  $(M_x - s_m \cdot h \leq \bar{M} \leq M_x + s_m \cdot h)$ , since the two assertions are formally equivalent;

(b) if I consistently assert that  $(M - s_m \cdot h \leq M_x \leq M + s_m \cdot h)$  the long run frequency of correct assertions I shall make will be  $P_t$  as defined in (ii) for all corresponding sample values  $M_x$  and  $s_m$  and for any single value the unknown parameter  $M$  may have;

(c) I may therefore associate an uncertainty safeguard  $P_f = (1 - P_t)$  with the rule:

for samples of which the relevant definite sample statistics are  $M_x$  and  $s_m$ , I shall always assert that  $(M_x - s_m \cdot h \leq \bar{M} \leq M_x + s_m \cdot h)$ .

In this formulation we do not make any self-contradictory assertion assigning a probability to the range of values in which a putatively unique parameter lies. Instead, we confine ourselves to prescribing a rule to which we can assign a long-run frequency of correspondence between our own statements and observed events. So far as it goes, this formulation is therefore entirely consistent with the Forward Look; but if we ask what the rule can do for us, we disclose an inadequacy. To set any preassigned precision level to our terminal statements about  $M$ ,

we need to know  $s_m$  as well as  $h$ , and we can do so only if we have first looked at the sample. Thus a rule stated in advance and followed consistently regardless of the outcome cannot guarantee that the outcome will be useful. It is either useless or inconsistent with the Forward Look.

Up to a point, Neyman concedes this in a recent discussion on a contribution by Stein. Stein has attempted to evade one horn of the dilemma by prescribing a two-stage procedure, which in effect provides the possibility of working within the framework of a normal universe for which  $\sigma$  has a fixed value; but Neyman's own candid comment (*Lectures and Conferences on Mathematical Statistics and Probability*, 1952) exonerates us from the need to refer to his sequential method more explicitly in this context:

Brilliant as his result is . . . its practical applications involve a new difficulty just as insuperable as that complained of by Berkson.\*

Wrighton's principle of *a priori adequacy* epitomises all that need be said from a formal viewpoint about the issue to which Berkson drew attention. Of itself the interval assigned by fixing an acceptable uncertainty safeguard by recourse to (ii) is not a proper rule of stochastic induction. To make it such we must take another step by assigning the size of sample sufficient to ensure in advance an interval length concordant with an acceptable level of precision. None the less, it is appropriate here to state the issue in another way, because the formulation of Confidence theory in the foregoing citation is equally inadequate when the relevant sampling distribution endorses the statement of a rule which conforms both to the formal requirements of a behaviourist outlook and the presumptive operational intention of a rule worth stating at all.

If we are to retain a useful place for the  $t$ -distribution as an instrument of interval estimation in the domain of representative scoring, we can do so only if we claim the right to rely on the evidence of a single sample when consistent with

\* J. Berkson first drew attention to the inconvenience of the fact that a confidence interval determined as above may exceed a precision level consistent with any conceivable practical requirements (*vide pp. 418-420 above*).



the terminal statement we deem to be acceptable. If we adopt the new orientation, we are then reserving the luxury to look at the sample before we decide whether to operate a rule which implicitly guarantees a particular set of terminal statements in advance. In mixed metaphor we have brought back the Backward Look by the back door. We are claiming the right to make the class of statements Confidence theory disclaims\* and—as we shall later see—the Fiducial argument condones.

Should any doubts remain about the propriety of this procedure, we may dispel them by soliciting the co-operation of the always serviceable Martian observer. We shall suppose that we set the problem of Model V (b) in Chapter 15 to a very large number of observers each with limited resources, which we may signify by allowing them to operate once with a sample of 100. We shall assume that the acceptable terminal statement places  $p$  in an interval of length  $i_c \leq 0.1$  as before with a risk of error  $P_f \leq 0.05$ . Now we have seen that a rule which guarantees this in all circumstances in its entirety calls for samples of 396 or more; but the terminal statements it permits places the value of  $p$  in an interval much shorter than 0.1 for observed values of  $p_0$  near the limits there assumed to be 0.1 — 0.9. Indeed, a corresponding rule will prescribe an interval of the required length for outsize values of  $p$  when  $r$  is much less than 396. As an exercise, the reader may apply the algebraic reasoning invoked on p. 413 to ascertain what values of  $p_0$  would place  $p$  in an interval  $i_c \leq 0.1$  when  $r = 100$  as we here assume; and to show how much the interval prescribed by consistent operation of the rule given on p. 415 would exceed the acceptable precision level for sample scores in the neighbourhood of 50 ( $p_0 \simeq 0.5$ ).

\* As the title of their paper *The Use of Confidence or Fiducial Limits* implies, Clopper and E. S. Pearson (1934) explicitly stated the most essential issue w.r.t. which Fisher's fiducial concept and Neyman's theory of Confidence part company long before their contemporaries recognised that a fundamental difference was at issue. Thus they say (*Biometrika*, 26):

... the percentage of wrong judgments differs according to the value of  $x$ , from 100 to 0. We cannot therefore say that for any specified value of  $x$  the probability that the confidence interval will include  $p$  is .95 or more. The probability must be associated with the whole belt, that is to say with the result of the continued application of a method of procedure to all values of  $x$  met with in our statistical experience.

If our observers take the stand that one is free to make statements when acceptable and otherwise refrain, they will therefore make statements to the effect that  $p$  lies within acceptable limits consistent only if they observe such outside values of  $p_0$ . From his backstage viewpoint, the Martian umpire would notice that each observer who recorded a score sum of e.g. 50 ( $p_0 = 0.5$ ) at a particular trial would omit to record the score, thus falsifying the balance sheet which correctly describes the outcome of their observations considered as a whole. If he understands the rule, and knows that  $p$  does not fall within the limits assigned by the recorded scores, he will not be fooled by the plea of poverty as an excuse for cheating. The assigned risk  $P_f \leq 0.05$  is relevant if and only if we follow the rule regardless of the outcome of any single trial.

This is why Wrighton's criterion of *a priori adequacy* is so indispensable to a statement of the problem of interval estimation wholly consistent with the Forward Look. Whereas Neyman is content to demand that the size of the sample shall ensure a form of statement consistent with an acceptable level ( $\alpha$ ) of risk, Wrighton insists that the form of statement also must conform to a criterion of acceptability. Otherwise, we may either end by saying nothing new or break the rule of the game by declining to follow it unless the result of doing so is congenial to our hopes and fears. While we must thus acknowledge a debt to Neyman for first stating what a rule of stochastic induction consistent with a behaviourist viewpoint implies at a verbal level, his successors who build on the foundations he laid will record that his algebraic formulation does not suffice to prescribe such a rule. It merely prescribes an important step in the process of devising one.

It happens that the numerical prescription for assigning fiducial limits within which  $M$  lies by recourse to the  $t$ -distribution is precisely the same as Neyman's numerical prescription for defining confidence limits in the same class of situations. It is mainly for this reason that many statisticians in the late 'thirties and early 'forties continued to regard the two theories of interval estimation as identical. At this level indeed the only difference between the two is a form of words. R. A. Fisher explicitly speaks of the probability that  $\mu$  (our  $M$ ) lies between

certain limits when  $\bar{x}$  (our  $M_x$ ) has a particular value. He inserts the epithet fiducial in the statement merely to emphasise his repudiation of inverse probability as Laplace propounds it, i.e. reliance on the scholium of Bayes. The reason why the  $t$ -distribution permits us to take this step is in his view:

"That the two quantities, the sum and the sum of the squares calculated from the data, together contain all the information supplied by the data concerning the mean and variance of the hypothetical normal curve. Statistics possessing this remarkable property are said to be *sufficient*, because no other can, in these cases, add anything to our information. The peculiarities presented by  $t$  which give it its unique value for this type of problem are:

- (i) Its distribution is known with exactitude, without any supplementary assumptions or approximations.
  - (ii) It is expressible in terms of the single unknown parameter  $\mu$  together with unknown statistics only.
  - (iii) The statistics involved in this expression are sufficient."
- (*Design of Experiments*, pp. 205-6, *First Edition*.)

Before examining what Fisher means by a sufficient statistic, let us examine his own verbal formulation of the problem discussed above. In Neyman's statement of the case, we do not speak of the probability that  $M$  lies within a particular range of values. In the domain of observable events this probability is in fact zero or unity and can have no other meaning. An essential step in the fiducial argument is the one we have indeed repudiated as inconsistent with a behaviourist viewpoint in our discussion of Model III on p. 397. Since the statement ( $M_x \leq M + h.s_m$ ) is formally equivalent to the statement ( $M \geq M_x - h.s_m$ ), we assume the right to state (iii) above in the form

$$P(M \geq M_x - h.s_m) = \int_{-\infty}^h F(t) . dt \quad . \quad . \quad (iv)$$

For observed values  $M_x$  and  $s_m$  referable to a sample we have observed, we may then construct a Fiducial Probability Distribution. The example cited on p. 418 will suffice to illustrate the procedure. There we observe  $M_x = 26.92$  and

$s_m = 6.64$  for a 12-fold sample. From the table of the  $t$ -integral we obtain for  $h = 2.2$ :

$$P(M_x \leq M - h.s_m) = \int_{-\infty}^{-h} F(t)dt = 0.025$$

$$P(M_x \leq M + h.s_m) = \int_{-\infty}^h F(t)dt = 0.975$$

$$\therefore P(M - h.s_m \leq M_x \leq M + h.s_m) = 0.950$$

By recourse to the table we obtain in the same way when  $M_x = 26.92$  and  $s_m = 6.64$ :

$h$	$P(M \geq M_x - h.s_m)$	$M_x - h.s_m$
$-3.11$	0.005	6.27
$-2.72$	0.010	8.86
$-2.20$	0.025	12.31
$-1.80$	0.050	14.97
$-0.70$	0.250	22.27
0.00	0.500	26.92
$+0.70$	0.750	31.57
$+1.80$	0.950	38.87
$+2.20$	0.975	41.53
$+2.70$	0.990	44.98
$+3.11$	0.995	47.57

For an observed sample in terms of relevant sample statistics (here  $M_x$  and  $s_m$ ), such a fiducial distribution of a parameter ( $p$ ) cites the probability that  $p$  will exceed an assigned value, and hence that it will lie within specified limits. If it is admissible to apply the rules of the classical calculus of probability in this context, it is then right and proper to state that

$$\begin{aligned} P(M_x - s_m.h \leq M \leq M_x + s_m.h) \\ = P(M \leq M_x + s_m.h) - P(M \leq M_x - s_m.h) \quad (v) \end{aligned}$$

The table of the  $t$ -integral gives:

$$P(M_x - 2.2.s_m \leq M \leq M_x + 2.2.s_m) = 0.975 - 0.025$$

When  $M_x = 26.92$  and  $s_m = 6.64$  the foregoing table of the fiducial distribution then gives

$$P(12.31 \leq M \leq 41.53) = 0.95$$

In Fisher's theory these are the 95 per cent fiducial limits of  $M$  for the observed sample whose definitive statistics are  $M_x = 26.92$  and  $s_m = 6.64$ . Numerically, they are identical with the confidence limits prescribed by Neyman's theory. The essential divergence is the step incorporated in (iv) above; and if we admit this step we might obtain the same result less circuitously by writing (ii) as

$$P(M_x - h.s_m \leq M \leq M_x) = \int_{-h}^h F(t).dt$$

If we do admit the step incorporated in (iv), and hence the possibility of constructing a fiducial probability distribution, we may proceed to use it in accordance with the rules of the classical calculus and the difference between the two formulations then assumes a more challenging aspect. This occurred when Fisher prescribed a recipe for interval estimation of the difference ( $M_b - M_a$ ) between the mean of two hypothetical universes A and B each normal but with different unknown parameters  $\sigma_a$  and  $\sigma_b$ . In the derivation, he takes a step which is certainly not consistent with Neyman's procedure which works within the framework of sampling distributions prescribed by the classical theory. A fiducial probability distribution is not a sampling distribution in the classical domain of events. The unknown parameter ( $p$ ) is merely a conceptual variable presumed to have a fixed value in the factual domain. In treating it as a variable, Fisher followed his own intuitions as his apologists have to concede, and the details of the controversy over the so-called Behrens test need not greatly concern us. Yates, who has the last word to date in defence of the Fiducial Argument, concedes (see *Appendix IV*, p. 506) that it is possible to justify it only if we invoke a new law of thought. Seemingly, this undertaking is not uncongenial to the converted. Others may harbour honest doubts about the difficulty of accommodating a customary minimum of intellectual rectitude

with the invocation of a new law of thought to rehabilitate an otherwise indefensible proposition in the absence of any ostensible additional advantages.\*

We are now in a position to recognise two circumstances which delayed recognition of a difference which now divides exponents of statistical theory into irreconcilable factions. One is that the  $t$ -distribution which provided the pivotal illustrative material for Neyman's original statement of his viewpoint cannot endorse the useful outcome of a rule stated before examining a sample and consistently pursued regardless of what structure any individual sample may have. The other is that the fiducial argument leads to numerical results inconsistent with the alternative approach only when we claim the freedom to operate with so-called fiducial probability distributions in accordance with the classical recipe for deriving the sampling distribution of a joint score of components with specifiable distributions referable to different universes of experience.

None the less, an essential difference expressly stated in a paper by Clopper and Pearson (1934) divides the two forms of statement at the outset. That a parameter definitive of a fixed framework of sampling has one value, that it therefore either does or does not lie within a particular range of values specified by an interval estimate, and that it is accordingly meaningless to assign any probability other than zero or unity to the statement that it does so in the domain of events, are niceties of verbal usage which we might well transgress, if a rule conceived in terms of the Forward Look could assign a probability to the truth of assertions based on particular observations. We might then regard the fiducial distribution of  $p$  referable to a particular observed value  $p_0$  as an elliptical way of describing

\* Kendall's summing-up of the Behrens test controversy is quotable, because the final remarks (*italics inserted*) have a much wider range of relevance to current statistical controversies:

So far as concerns problems of estimation, the Behrens test is accurate both in fiducial theory and in the theory of probability propounded by Jeffreys. But the test does not hold in the theory of confidence intervals. In fact the latter fails to provide an exact solution of the problem. . . . Fisher has criticised Confidence intervals on the grounds that they do not give an answer to what is admittedly an important question; but *it appears possible to maintain consistently that some questions may not have an answer.*

the probability of making correct assertions about  $p$  when our source of information is indeed  $p_0$ . What Neyman calls a rule of inductive behaviour does not in fact justify statements of this sort. The probability which we denote as the uncertainty safeguard of the rule pertains only to the entire class of statements subsumed thereby when we operate the rule consistently in its entirety.

From this point of view, the concept of sufficiency which occupies so prominent a place in the fiducial argument is irrelevant to the theory of interval estimation explored in the last two chapters. An air of mystery envelops the sufficiency concept, because current standard textbooks commonly illustrate it by reference to continuous distributions and hence enlist manipulative skill for dealing with multiple integrals beyond the range of readers who are not trained mathematicians. Actually, the essentials are easy to grasp in the terrain of classical models, such as the following. We assume two universes from which we may sample:

A. A full card pack for which we assign zero score if a card is black and unit score if a card is red;

B. A full card pack for which we assign scores of 0, 1, 2, 3 respectively to cards specified by suit as clubs, diamonds, hearts and spades.

We shall now consider the result of extracting *successively* from each universe 5 cards without replacement. We may then denote the component unit scores as  $x_1, x_2 \dots x_5$  and the probability that the sample is a given sequence as  $P_s(x_1, x_2 \dots x_5)$ . Of many ways in which we may score the 5-fold trial one is to record the mean value of  $x$ , and we shall assume that the mean score  $x_{m.5}$  of the 5-fold trial is 0.6. If the sample comes from universe A, we then know that it consists of 3 red and 2 black cards. Whence from (iv) of p. 41 we may write

$$P_s(x_1, x_2 \dots x_5) = 26^{(3)} \cdot 26^{(2)} \div 52^{(5)}$$

Thus the mean score suffices to specify the probability of meeting a sample of relevant specified make-up. More generally for a 2-class universe of  $s$  objects to which we assign

a score  $b$ , the mean score of the  $r$ -fold non-replacement sample suffices to specify the probability of getting the score sequence  $x_1, x_2, x_3 \dots x_r$ , since any unit score  $x_u$  in the sequence must have the value  $a$  or  $b$  and  $x_{m,r}$  therefore suffices to evaluate  $u$  and  $v = (r - u)$  in the expression:

$$P_s(x_1, x_2, x_3 \dots x_r) = (s^{(u)} \cdot f^{(v)}) \div n^{(r)}$$

In this sense we may say that the statistic  $x_{m,r}$  contains all the relevant information a sample from a 2-class universe can contain; but this is not true of the mean score referable to the alternative 4-class universe B above. If  $r = 5$  all we may infer from the information  $x_{m,5} = 0.6$  is that the composition of the sample may be any of the following:

<i>Unit Trials</i>	<i>Probability of the observed sequence</i>
4(0) + 1(3)	$13^{(4)} \cdot 13 \div 52^{(5)}$
3(0) + 1(1) + 1(2)	$13^{(3)} \cdot 13^2 \div 52^{(5)}$
2(0) + 3(1)	$13^{(2)} \cdot 13^3 \div 52^{(5)}$

As we see from the above, each unique *combination* of unit trial scores here assigns a different probability to the observed event, i.e. particular combination in a particular sequence.

This example exhibits the meaning we may attach to the terms when we say that  $x_{m,r}$  is a *sufficient* statistic when we sample in universe A and an *insufficient* statistic when we sample in universe B; and the formal distinction is easy to infer from it. For samples from universe A, the probability assignable to  $x_{m,5} = 0.6$  is:

$$P_{s,m}(x_1, x_2 \dots x_5) = \frac{3! 2!}{5!}$$

If  $P_m$  is the probability of getting a unique score combination:

$$\begin{aligned} P_s(x_1, x_2 \dots x_5) &= P_{s,m}(x_1, x_2 \dots x_5) \cdot P_m \\ &= \frac{3! 2!}{5!} \cdot \frac{5!}{3! 2!} \cdot 13^{(2)} 13^{(3)} \div 52^{(5)} \\ &= 13^{(2)} \cdot 13^{(3)} \div 52^{(5)} \end{aligned}$$



Thus we may here split the expression for the probability of extracting a sample completely specified as a unique ordered sequence into 2 factors, one being the probability that the sufficient statistic will have its observed value, the other being the conditional probability of the observed event if the sufficient statistic has this value. Now we cannot do this with the insufficient statistic  $x_{m,r}$  referable to sampling in universe B. We then have for  $r = 5$  and  $x_{m,5} = 0.6$

$$52^{(5)}.P_m = \frac{5!}{4!1!} 13^{(4)}13 + \frac{5!}{3!1!1!} 13^{(3)}13^2 + \frac{5!}{2!3!} 13^{(2)}13^3$$

If it happens that the sample consists of 4 clubs and 1 spade, we may write

$$52^{(5)}.P_s(x_1, x_2 \dots x_5) = 13^{(4)}13 \text{ and } P_{s,m}(x_1, x_2 \dots x_5) = \frac{4!1!}{5!}$$

Whence it is clearly false to write as above:

$$P_s(x_1, x_2 \dots x_5) = P_{s,m}(x_1, x_2 \dots x_5).P_m$$

The possibility of factorisation in this way does indeed furnish the formal conditions which define whether a sample statistic is sufficient; but the status of the concept of sufficiency in the theory of interval estimation is amenable to discussion at a less formal level. The important difference between  $x_{m,r}$  as a sample statistic of universe A and  $x_{m,r}$  as a sample statistic of universe B resides in the fact that the former alone embodies every relevant particular referable to the individual sample. If then we claim the right to assign a probability to statements about a source on the evidence which a sample of specified make-up supplies, we shall employ statistics which summarise all the evidence a sample can indeed supply, whence we shall invoke only sampling distributions referable to sufficient statistics. If we claim no more than the right to associate a probability to correct assertion within the framework of consistent adherence to a rule stated in advance, there is no obvious reason why we should conform to any such restriction.

The controversy over the Behrens distribution and the search for sufficient sampling statistics has provoked an

extensive literature at a high level of mathematical sophistication, but the *casus belli*, like the *t*-distribution dilemma disclosed in our earlier discussion of Model VI (p. 416) involves logical issues which are not difficult to grasp if we keep our feet on the solid ground of situations for which we can prescribe *discrete* distributions as in Chapter 14 and 15. The dilemma raised in the Model VI set-up does not arise, if we endorse Wrighton's principle of *a priori adequacy* as an inescapable obligation to those who adopt the Forward Look. The *t*-distribution relies exclusively on sufficient statistics; but it cannot provide a sufficient basis for the interval estimation of the mean of a normal distribution in terms consistent with a behaviourist viewpoint, since the rule prescribed admits of no restriction to the set of terminal statements it endorses. Meanwhile, controversy continues in a progressively more rarefied atmosphere of symbolic conventions remote from the real world. So far the outcome has been to provide pure mathematics with a new crop of problems. That the illumination conferred will prove to be proportionate to the output of heat generated seems less certain. It is the writer's view that further progress awaits a vigorous restatement of first principles, and that such restatement will reveal both the need for refinement of the original terms of reference of Neyman's theory and the impossibility of accommodating the Fiducial Argument to an interpretation of the terms of reference of a calculus of probabilities in the domain of observable occurrences.

Kendall (*op. cit.* Vol. II, p. 90) clearly recognises the difficulty of accommodating Fisher's formulation with the behaviourist viewpoint, when he states

... Fisher considers the distribution of values of  $\theta$  for which  $t$  can be regarded as a representative estimate-representative, that is to say, in the sense that it could have arisen by random sampling from the population specified by  $\theta$ . As pointed out above, this does not mean that we are regarding the true value of  $\theta$  as a member of an existing population. Rather, we are considering the possible values of  $\theta$  and attaching to each value a measure of our confidence in it, based on the probability that it could have given rise to the observed  $t$ .

If I interpret him correctly, Fisher would regard a fiducial dis-

tribution as a frequency-distribution. This implies that  $\theta$  is regarded as a random variable. It appears to me, however, that it is not a random variable in the ordinary sense of the frequency theory of probability, in which values of  $\theta$  either are or can be generated by an actual sampling process. We can never test whether the fiducial distribution holds in the frequency sense by drawing a number of values and comparing observation with theory. Nor, in calculating fiducial limits of the type  $\theta = t + h(\alpha)$ , do we imply that the proportion of cases for which  $\theta \leq t + h$  is true will be  $\alpha$  in the long run.

We have sufficiently acquainted ourselves with the inadequacy of Neyman's original statement of the problem if we accept the principle of *a priori* adequacy. Wrighton specifies a second restriction likely to call for more drastic innovations of procedure if it commends itself to our good judgment. This is the *principle of the minimal set*. The principle of *a priori* adequacy signifies that the rule must guarantee the possibility of making terminal statements which are acceptable at the lowest level in the sense that we must always be in a position to say more than the outcome of non-stochastic reasoning in the same class of situations. The principle of the minimal set imposes a three-fold limitation. If less compelling than the principle of *a priori* adequacy, it does at least link the theory of estimation more closely to the real world. A set of acceptable statements is minimal in Wrighton's sense, if it conforms to three requirements:

- (i) *No terminal statement may be null in the sense that it implies reservation of judgment.*
- (ii) *No terminal statement may be inconsistent with background knowledge of prior possibilities.*
- (iii) *No terminal statement may imply a higher level of precision than the rule in its entirety can endorse.*

The null condition is trivial in the domain of interval estimation and merely prescribes that the rule of stochastic induction should be comprehensive in the sense defined on p. 346. Some would deny that it is a logical necessity, but the practical inconvenience of relinquishing the requirement in the domain of the alternative test procedure should be sufficiently clear in

the light of the discussion on pp. 359–362 of Chapter 15. Disregard of (ii) above, viz. the requirement that no terminal statement should necessarily be false, might likewise seem to be logically consistent with what we claim for the rule as a whole; but it is inconsistent with the principle of design implicit in the interpretation of a confidence chart like that of Fig. 3 as a visualisation of the hypothesis-observation grid in the two-way continuum. This emerges clearly from our discussion of Models V (a) and V (b) and IV in Chapter 15.

The incorporation of (ii) as a fifth (see p. 432) canon of stochastic induction in an adequate restatement of the theory of confidence intervals will remove a widely felt objection to Neyman's original theory and one which has prompted the criticism that Neyman dispenses with the prior probabilities of Bayes by an ingenious trick. We have seen sufficient reason to exonerate Neyman from this charge, since Bayes's prior probabilities are assignable to hypotheses in a meaningful sense only in the domain of what von Mises speaks of (p. 138) an experiment carried out in two stages, i.e. when each of the hypotheses of the comprehensive set is referable to a real population at risk. The form of the objection so stated thus arises from failure to distinguish between the *prior probabilities* of Bayes and the *prior possibilities* which delimit the comprehensive set of hypotheses, whence also the rightly conceived boundaries of the hypothesis-observation grid. None the less, the need for a realistic formulation of any rule of stochastic induction with due regard to the latter remains, as Kendall (*Biometrika* XXXVI, 1949) recognises:

Suppose I assume that a sampling process is such as to reproduce a binomial distribution—there is a good deal of evidence for this in the case of births. I observe a value of 0.60 as the ratio of male to total births in a sample of 10,000. The theory of confidence intervals says that I may assert that the proportion  $p$  lies between 0.59 and 0.61 with the probability that, if I make this type of assertion systematically in all similar cases, I shall be about 95 per cent right in the long run. But I do not then make such an assertion because I know too much about birth-rates to believe any such thing. The theory of confidence intervals gives no place to prior knowledge of the situation. How, then, can that theory provide a guide to conduct in making decisions?

The third requirement embodied in Wrighton's principle of a *minimal* set of acceptable terminal statements confers a special relevance on the italicised epithet. Its importance emerges only when the terminal statement specifies an interval  $(p_a - p_b) = i_{p,0}$ . In some model situations, e.g. that of p. 359 in Chapter 15, a rule which endorses the statement  $M_a = k$  at an acceptable uncertainty level  $P_f \leq \epsilon$  for a particular observation endorses the same identity for all situations; but this need not be so. It is not true of the Model III at the end of Chapter 14; nor is it true of Model V (b) in Chapter 16. There we saw that if  $i_c = k$  for  $p_0 = 0.5$ ,  $i_c < k$  for all other values of  $p_0$  in the admissible range.

Now it is not necessarily inconsistent with the operation of the rule in its entirety to allow that different observations may prescribe different values  $k_{p,0} < k$ , the requisite precision level which determines the requisite size ( $r$ ) of sample consistent with the assigned uncertainty safeguard  $P_f \leq \epsilon$ ; and we have adopted this procedure in the discussion of Model V (a) in conformity with Neyman's formulation. On the other hand, there are good reasons for relinquishing the obligation to do so, and if we do dispense with it we embrace certain advantages, in particular the possibility of handling otherwise intractable situations. We are then free to formulate rules of interval estimation, which are not identical with those prescribed by Neyman's formal definitions, and rules which also incorporate (ii) above.

Wrighton's argument is as follows. Presumptively our end in view, when we operate within the framework of the hypothesis-observation grid either in non-stochastic or stochastic terms, is to decide whether a statement is admissible within the corpus of knowledge; and our precision level ( $k$ ), being in this context our criterion of acceptability, defines the eligibility of a terminal statement to admission within the corpus of knowledge. If we take the concept of such a corpus of knowledge seriously, the admission of terminal statements to the effect  $k_{p,0} < k$  when  $p_0$  has particular values other than the value which endorses  $k_{p,0} = k$  would seem to imply a *second* corpus of knowledge with which we associate more exacting standards of precision than the operation of the rule in its

entirety can guarantee. If so, the experimental design violates the principle of *a priori adequacy* relative thereto. The argument so stated is subtle. In more simple terms, we may state it interrogatively thus: if I say that I ask no more of my rule in its entirety than the right to assert  $k_{p,0} = k$ , have I really gained anything by reserving the right to say  $k_{p,0} < k$  in particular situations?

Though this does not call for restatement of the method of determining the interval in the domain of Model VI (p. 416) of Chapter 17, it will lead us to a different procedure in the situation of Model V (*b*). In the case which arises when the binomial parameter  $p$  of Model IV may admissibly have any value in the range 0 — 1.0, two parallel straight lines like those of Fig. 4 in our treatment of Model VI enclosing the lozenge shaped region of Fig. 3 will define all the intervals corresponding to our set of acceptable terminal statements. Consequently, the numerical computation of confidence limits prescribed by the adoption of the Principle of the Minimal Set will not be as for Neyman's theory of confidence intervals.

Wrighton's restatement differs from Neyman's in another way. As we have consistently maintained, Neyman insists that a rule of induction referable to the terms of reference of the classical theory is a rule stated in advance. Thus the calculus of judgments cannot prescribe how we should weigh the evidence which a single sample supplies. It can merely endorse a procedure for weighing the way in which we weigh the evidence. None the less, it would seem that Neyman still experiences a nostalgia for retrospective judgments in so far as he seemingly condones the practice of examining previously accumulated data as if they are accumulating in the course of an experiment designed in accordance with what he calls a rule of inductive behaviour. At first sight, Wrighton's principle of the minimal set might seem to make this licence more defensible, since it excludes our right to make statements prescribing a precision level more refined than the size of sample can endorse. Wrighton repudiates this deviation from the Forward Look. Any such appeal to the restriction the principle imposes is illusory, if we consistently interpret the true terms of reference of a rule stated in advance. As Clopper

and Pearson remind us, we relinquish more than the certainty of being right when we operate a rule of stochastic induction with an assignable risk of error. We confer on the totality of our assertions a greater precision, but we can assign no acceptable risk to the possibility that any individual assertion is false. We cannot admit the right to examine isolated samples of past experience as if we were performing experiments in accordance with a rule which disclaims any title to evaluate the verdict we may pass on any single one of them.

A communication (see p. 24) of Anscombe on the Analysis of Variance was the occasion for a discussion which forced the current controversy over the credentials of statistical inference into the open. Anscombe was content to point out that the null hypotheses of the test battery are factually ambiguous. Wald's restatement of test procedure as prescribed by Neyman and E. S. Pearson seemingly deprives the battery of any intelligible rationale. What then remains of the technique we now identify with statistical design of experiments? In one of the few thought-provoking discussions of the Analysis of Variance, Churchill Eisenhart has advanced the view that its main usefulness lies in the domain of interval estimation. If we adopt Wrighton's restatement of Neyman's Confidence theory, we must abandon the hope of rehabilitating its credentials as such. What I have said, and what Neyman himself now concedes, concerning the limitations of the Gosset sample distribution as a device for assigning an interval estimate to the mean proscribes the Analysis of Variance as a procedure for so doing.

## CHAPTER NINETEEN

### EPILOGUE

IN THIS BOOK our concern has been to trace disputable claims of statistical theory to their sources. We have seen that the calculus on which current statistical theory relies took shape when the only practical preoccupation of the pioneers was a regimen for gambling in games of chance. With the rise of life insurance the emergence of a new *motif* signalises two innovations equally alien to any such undertaking. In effect, the calculus of probability henceforth claimed as its province situations in which the toy changes during the course of the game and the player is at liberty to change his bet as the game progresses. In the background of every contemporary issue, we may detect an uneasy but not fully articulate recognition that the calculus has no necessary relevance to situations of either sort. Having striven to make this twofold dilemma explicit it remains for me to summarise the outcome of the foregoing discussion as I now see it. I propose to do so by asserting three theses.

1. Unless we rely on axioms which are not susceptible of proof, we must concede that a calculus of probability is relevant to the real world; (a) only in so far as it specifies frequencies of observable occurrences in an indefinitely protracted sequence of trials; (b) only if also such occurrences ( $e_1$ ,  $e_2$ ,  $e_3$ , etc.) collectively constitute a sequence wholly devoid of order.

2. If we take the view last stated, it is improper to speak of the probability that a verdict on the outcome of a single trial is true. We can speak with propriety only of the frequency of correct assertion in an unending sequence of trials, and then only if we adhere consistently to the same rule. Since the theory of probability conceived in the foregoing terms does not sanction the right to vary the rules of the game in accordance with the player's luck, we then shoulder the obligation of *stating in advance the rule in its entirety*.

3. If we concede 1 (a) above, any proper terms of reference we may claim for a stochastic calculus of error, for a stochastic



## EPILOGUE

calculus of exploration or for a stochastic calculus of judgments restricts their legitimate use to situations of which we can predicate a fixed framework of repetition.

1. *Unless we rely on axioms which are not susceptible of proof, we must concede that a calculus of probability is relevant to the real world: (a) only in so far as it specifies frequencies of observable occurrences in an indefinitely protracted sequence of trials: (b) only if also such occurrences collectively constitute a sequence wholly devoid of order.*

At the outset we may dismiss any objection to an axiomatic approach when we invoke the calculus of probability to construct hypotheses subsumed by the subject-matter of Chapters Twelve and Thirteen. From the viewpoint of scientific method, the hypothesis that particles of matter behave in accordance with a stochastic model is on all fours with the hypothesis that matter is particulate. We accept or reject such hypotheses because they lead us to conclusions which are or are not *independently* verifiable; and a single accredited experiment of which the outcome is conclusively inconsistent with the axioms suffices to compromise their further usefulness irremediably. Thus none of the foregoing theses impinge on the credentials of what I have called a stochastic Calculus of Aggregates. Contrariwise, the admitted usefulness of stochastic models in contemporary physics and genetics has no bearing on the claims of the theory of probability in situations which offer us no opportunity for testing the validity of our initial assumptions in the light of their practical consequences.

If we invoke the calculus of probability in such situations, my first thesis signifies that we implicitly accept the responsibility of defining circumstances in which we can either identify randomness as a property of a system of occurrences or conduct our observations within a framework of randomisation which we ourselves impose. Now the only domain of action in which it has hitherto been possible both to test the credentials of the calculus and to record an outcome favourable to its claims is the classical domain of games of chance. We may justifiably invoke it in other domains of action, only if we are able to specify what is common to all

the procedures subsumed by games of chance. In all such games, we recognise an inextricably three-fold relation between an agent, an apparatus and a randomising procedure. We have seen no sufficient grounds for predicating randomness as something which the apparatus generates except in so far as an agent conforms to the programme of active interference subsumed by the randomising procedure; and all we have been able to say with confidence about the latter is that the task it sets transcends the limits of sensory discrimination of the agent.

It is thus clear that the stochastic Calculus of Error stands on a firmer foundation than its offspring which I have referred to elsewhere as the stochastic Calculus of Exploration. Though the latter derives its formal symbolic outfit with only trivial refinements from the former, it relegates to Nature the responsibility which the player undertakes in the classical situation. Contrariwise, the player takes his proper place in a stochastic calculus of error as the observer. In the same idiom, *accidental* error is precisely equivalent to a score component associated with that part of the player's assigned task beyond possibility of fulfilment in virtue of the limits of sensory discrimination. *Mutatis mutandis*, we may say the same about certain statistical procedures employed in production engineering.

As it bears on the status of what I have called a stochastic Calculus of Judgments, my first thesis raises an issue not as yet examined in these pages. Throughout the foregoing discussion, we have assumed a system of randomwise sampling without discussing what circumstances justify the identification of such a system with situations in which we apply a test procedure or undertake an interval estimation. The contemporary attitude to this task is puzzling. In one context we read that we can legitimately regard any sample (*vide*, p. 489) as a random sample from an infinite hypothetical population, and in another context one and the same author will advocate recourse to random numbers to ensure truly randomwise sampling. If we reject the proposition that mere ignorance is a sufficient guarantee of random selection and decline to assume that the Divine Dealer shuffles the *billets* in Nature's urn before each draw, we must indeed accept the obligation to employ some

randomising device when we apply procedure discussed in the last four chapters.

Our examination of the credentials of a calculus of judgments will therefore be incomplete unless we seek an answer to the question: how can we guarantee that a system of sampling is truly random? All statisticians hold that artificial randomness is attainable. Most of them subscribe to the use of *random numbers* as an appropriate device for attaining the end in view; but few consumers are alert to the implications of their use. Since some readers may be unfamiliar with a table of such numbers, a few preliminary remarks on the prescribed usage will not here be out of place. A table of random numbers is a table of  $c$  cyphers (0-9) in each of  $r$  rows selected in various ways (*vide infra*). If the table is extensive enough, we may expect to meet any of the first hundred integers (labelled 00-99) in the first two columns, any one of the first thousand (labelled 000-999) in the first three, and so on. The way in which we use the table depends on the size of the total sample ( $N$ ) and the number of groups into which we seek to divide it ostensibly randomwise.

To illustrate the prescribed procedure, we may here suppose that  $N = 96$  and that we wish to divide an  $n$ -fold sample of persons into three subsamples of 32 each. We may then decide to label the three groups by the numbers:

00-32 A : 33-65 B : 66-99 C

To allocate our 96 persons to the groups, we first label the individuals as  $P_1, P_2, P_3, \dots P_n$  quite arbitrarily. Having done so we say that the random number for  $P_1$  is the first pair of cyphers in the first column, that of  $P_2$  is the second, and so on. The following example will make this clear. As we go from top to bottom, the first two columns in one set of tables read: 03, 97, 16, 12, 55, 16, 84, 63. . . . Thus we award to the person ( $P_8$ ) whose rank order is 8, the number 63, which places him in group B. In proceeding thus, we shall not expect the groups to accumulate their quota simultaneously. Thus the first 24 rows of the first two columns of the table cited would lead to the allocation of 11 persons to A, 7 to B and 6 to C.

There will come a stage when we have allocated 32 to one group, let us say A. We then neglect any table entry which assigns a number in the range 00–32 and pass on to the next, till we have filled up the quota for a second group, let us say B. In that event we allocate all the rest to C.

The validity of any such procedure raises several issues, including whether the make-up of the table sufficiently guarantees an Irregular Kollektiv of cypher sequences. Kendall cites three methods of constructing one;

(i) *Use of Mathematical Tables*. Fisher and Yates derived theirs from the 15th to the 19th digits of A. J. Thompson's table of logarithms;

(ii) *Use of Census Figures*. The assumption is that the final digits of large numbers taken from demographic data in no preassigned order will turn up randomwise. Tippett adopted this plan.

(iii) *Mechanical*. By definition, an ideal lottery wheel with 100 sectors labelled 00–99 will generate a random sequence of paired cyphers. On this assumption, Kendall and Babington Smith have tabulated such number sequences by specially constructed machines.

If we ask why successive final digits or pairs, etc., of digits in a table of logarithms should constitute a random sequence, one answer we may get appeals to the fact that a logarithm is a transcendental number; but the fact that such a number cannot be the solution of an algebraic equation in the current sense of the term does not obviously imply that the sequence of cyphers by which we represent it constitutes an Irregular Kollektiv as von Mises uses the term. Moreover, the method is open to a formidable objection of another sort. Kendall expresses it thus:

Here again, however, the use of such tables requires care—they may have been compiled by an observer with number preferences, and some rounding up may have taken place.

More explicitly, Kendall comments on the tables of Fisher and Yates in the following terms:

## EPILOGUE

These numbers were obtained from the 15th–19th digits in A. J. Thompson's tables of logarithms and were subsequently adjusted, it having been found that there were too many sixes.

The procedure followed by Tippet depends on the appeal to insufficient reason. We are then back to the axiom that blindfold selection from Nature's urn is *ipso facto* randomwise. If we accept it, we have exonerated ourselves from the obligation to enlist an artificial randomising device. If we reject it because we have insufficient reason for believing it to be true, there is nothing more to be said in favour of Tippet's method. It is justifiable only if also redundant.

The third method has a more attractive aspect, but Kendall himself concedes:

Thus, it is to be expected that in a table of Random Sampling Numbers there will occur patches which are not suitable for use by themselves. The unusual must be given a chance of occurring in its due proportion, however small. Kendall and Babington Smith attempted to deal with this problem by indicating the portions of their table (5 thousands out of 100) which it would be better to avoid in sampling experiments requiring fewer than 1,000 digits.

The mechanical procedure is attractive because we are at home in a familiar situation. We entrust the making of the table to a lottery, and we can all agree that a lottery situation is the type of situation in which it is indeed possible to generate a sequence as nearly random as we connive in the conduct of tests of the credentials of the calculus of probability hitherto undertaken. With the reservation last stated, we may therefore say that a table of  $r$  numbers consisting of  $c$  cyphers compiled in accordance with the outcome of  $r$  spins of a wheel with  $c$  sectors should record *one*  $r$ -fold sample in the unending random sequence. None the less, there remains a doubt about the relevance of the assertion to the task of allocating our  $N$  individuals randomwise to three groups, A, B and C? Surely, the task so stated implies a procedure which will justify itself only by continual repetition. If so, would not the correct course be to operate the lottery procedure anew whenever we make such an allocation?

The issue last raised focuses attention on an exceptionable

feature of *any* method of allocation which relies on a table of so-called random numbers; but the seemingly unexceptionable alternative suggested in the final query of the last paragraph evades a difficulty. When we use such a table our aim is not merely to allocate individuals in conformity with a random sequence. The end in view is also to ensure that every one of  $N$  individuals has an equal chance of allocation to one of the subsamples and that any assemblage of  $n$  individuals we may allocate to a particular subsample has the same chance of allocation thereto as any other assemblage of  $n$  individuals. Thus the task of prescribing randomwise allocation and that of recognising situations in which the calculus is operative bring us face to face with the same dilemma. If the behaviour of an ordinary cubical die or of a lottery wheel is inconsistent with the assumption of a rectangular distribution of unit trial score values, are we to conclude that: (a) the behaviour of the system fails to conform to the requirements of the Irregular Kollektiv; (b) though the score sequence is truly random, there is a bias in the sense that our specification of its definitive sampling parameters is erroneous?

If the series of the event is truly random, we require information about the outcome of an infinite sequence of unit trials before we can decide what the bias is. Experience of a finite sequence can merely endorse the upper limit of uncertainty to a statement concerning the limits between which the numerical value of a parameter must lie. To be sure, we can make both the limits and the associated uncertainty safeguard trivial if we conduct a sufficiently prolonged sequence of trials; but we cannot resolve the dilemma last stated by appeal to experience *a posteriori*. Any precise statement about limits endorsed by the calculus of probability is valid only if we can first assure ourselves that the series of the event is truly random.

In this context, the word *bias* as commonly used is a pitfall. We have no sufficient reason for assuming that any formal definition of probability, e.g. in terms of set theory (p. 49), embraces the specification of the true numerical values of parameters definitive of a sampling distribution. Nor have we sufficient reason for assuming that the exclusive source of any

departure from values assigned *a priori* is some mechanical defect of the apparatus, in contradistinction to minor variations of procedure consistent with the *explicit* programme of instructions. If we say that the probability ( $p$ ) of tossing a head lies in the range  $\frac{1}{2} \pm b$ , we cannot therefore legitimately imply assent to the proposition that  $p$  has a unique limiting value fixed by the agent's instructions and the construction of the penny.

If we could confidently assume that the programme of instructions guarantees a truly random sequence in which  $p$  has a unique limiting value, it would be possible to score the results of trials to compensate for *bias* in the customary and naïve sense of the term, e.g. by alternately scoring heads as 0 or 1 and tails likewise. We should then have a random sequence with equiprobability of scoring unity or zero. We might accomplish the end in view in many other ways. For heuristic reasons, it will be helpful to take a backstage view of a particular system of compensatory scoring on this assumption. Our rule will be: if the number of tails in an  $r$ -fold trial is odd, score the result as 0, otherwise as 1. For  $r = 2$  we shall then set out our results thus:

<i>Result</i>	HH	HT or TH	TT
<i>Score</i>	1	0	1
<i>Probability</i>	$p^2$	$2p(1 - p)$	$(1 - p)^2$

Similarly for  $r = 3$  we should have:

HHH	HHT, HTH, THH	TTH, THT, HTT	TTT
1	0	1	0
$p^3$	$3p^2(1 - p)$	$3p(1 - p)^2$	$(1 - p)^3$

On the naïve view under consideration, we may assume a fixed proportionate bias  $b$  from the ideal value  $\frac{1}{2}$ , such that  $p = \frac{1}{2}(1 + b)$ . Thus for  $r = 2$ , the probability of scoring 1 is  $\frac{1}{2}(1 + b^2)$ , for  $r = 3$  the probability of scoring 1 is  $\frac{1}{2}(1 + b^3)$ .

If the proportionate bias thus naïvely defined is  $0.2$ , so that  $p = 0.6$ , it follows that the probability of scoring  $1$  in a single trial is  $0.6$ , in a double trial  $0.52$ , in a 3-fold trial  $0.508$ , and so on. More generally, the probability of scoring  $1$  in an  $r$ -fold trial would be  $\frac{1}{2}(1 + b^r)$ . On the assumption that  $b$  is fixed and does not exceed an agreed figure, we should therefore be able to make the probability of scoring  $0$  in an  $r$ -fold trial as small as need be by prescribing a sufficiently large value of  $r$ .

We have here assumed that the outcome of successive individual tosses is a truly random sequence. Now the assumption that it will be so involves the agent's intervention and the programme of instructions as well as the construction of the penny. Alas we have little reason to believe that it is possible to frame a programme which does not admit minor variations, any one of which will lead to a unique value of  $p$  in the long run if the agent follows the same plan at every trial. Some of them might favour heads, others tails uppermost at a toss, and if aware of the outcome, the agent might use his or her knowledge to vary the programme in an orderly way. In what follows, therefore, we shall no longer assume that the penny has a fixed bias as writers on the theory of probability commonly use the term. We shall also need to examine what are the consequences of withholding knowledge of the outcome from the agent.

All we shall then be able to say about the limits of bias will concede that the agent may be free to act both in a way which favours heads and in a way which favours tails. We suppose that one way of consistently behaving leads in the long run to an upper value  $p_1 = \frac{1}{2}(1 + b_1)$  and another to a lowest value  $p_2 = \frac{1}{2}(1 - b_2)$ . It will not affect the ensuing argument if we put  $b_1 = b = b_2$  for the outside limits. Accordingly, we may write  $p_1 = (1 - p_2)$  and  $p_2 = (1 - p_1)$ . We are then ready to appreciate in outline the rationale of a method of randomisation with equiprobability advanced by Wrighton (in the press). As a method of allocation, it is not open to the objection already advanced against the use of a table. It invokes the randomising procedure anew for every allocation of a group, and the randomising procedure itself has three essential innovations:



## EPILOGUE

(a) a blindfold agent tosses a penny in accordance with precise instructions with reference to the way in which he must toss it;

(b) an umpire records the score at each toss without divulging it to the agent;

(c) the scoring of the composite ( $r$ -fold) trial which determines the allocation of each individual follows the prescription already explored, viz. score 0 if the number of tails in the  $r$ -fold trial is odd, otherwise score 1.

The importance of withholding knowledge of the outcome of the toss from the agent will be clear, if we consider the 2-fold composite trial. For brevity we may put  $p_1 = \frac{1}{2}(1 + b)$  and  $p_2 = \frac{1}{2}(1 - b)$  for the definitive parameters of the two most-biased ways of interpreting the programme of instructions. If the agent does not know the outcome, he or she may operate either plan consistently or vary them without reference thereto; and we may prescribe possible results within limits set by regularly alternating or consistently adopting one or the other. We may then lay them out thus:

	HH	HT or TH	TT
	1	0	1
Plan 1	$p_1^2$	$2p_1(1 - p_1)$	$(1 - p_1)^2$
Plan 2	$p_2^2$	$2p_2(1 - p_2)$	$(1 - p_2)^2$
Alternating	$p_1p_2$	$p_1^2 + p_2^2$	$p_2p_1$

If the agent follows either plan consistently, the result will be the same. The probability of scoring 1 will be  $\frac{1}{2}(1 + b^2)$ . If the two plans alternate, it will be  $\frac{1}{2}(1 - b^2)$ . Thus the limits between which probability of scoring 1 must lie will be  $\frac{1}{2}(1 \pm b^2)$ . More generally for an  $r$ -fold trial, this will be  $\frac{1}{2}(1 \pm b^r)$ . Thus the system of scoring ensures that the probabilities of scoring 0 or 1 approach equality as  $r$  increases.

Now this is not necessarily so, if knowledge of the result of the toss empowers the agent to vary the plan of action in an *orderly* way. For instance, the agent would then be free to

operate plan 1 consistently if the result of the first toss of the 2-fold trial proved to be a head, but to operate plan 1 at the first and plan 2 at the second if the first proved to be a tail. Our schema thus becomes:

HH	HT	TH	TT
1	0	0	1
$p_1^2$	$p_1(1 - p_1)$	$p_2(1 - p_2)$	$(1 - p_1)(1 - p_2)$

On this showing the probability of scoring 1 would be  $\frac{1}{2}(1 + b)$ , i.e. the departure from equiprobability would be as great as if the agent scored the result in the usual way.

Without here defining our criterion for deciding on the number ( $r$ ) of tosses in the composite trial scored as 0 or 1 with probability as nearly 0.5 as we care to make it, we may now outline the method of allocation. To do so, we may speak of any such composite  $r$ -fold trial based on the primary sequence of tosses as a unit secondary trial, and of any  $k$  successive unit secondary trials as a  $k$ -fold secondary trial. The outcome of any one of these will be one of  $2^k$  different sequences of unit scores 0 or 1. We may imagine that each of an assemblage of  $2^k$  individuals receives a different ticket bearing one or other of these sequences. If we wish to allocate him to two groups of equal size, we may place any individual in Group A if the result of a  $k$ -fold secondary trial tallies with his ticket, continuing the process till Group A holds half the assemblage of  $2^k$  individuals. The remainder will then constitute Group B. If the size of the assemblage is  $N$ , it will not commonly be possible to fix  $k$  so that  $N = 2^k$ . All that matters is that each individual holds a unique ticket recording a score sequence which has approximately the same probability of turning up as any other. Thus  $2^k$  must be at least as great as  $N$  and our criterion for choosing  $k$  will be  $2^{k-1} < N \leq 2^k$ .

We have still to define how to fix  $r$ . If the end in view is that the ticket held by every individual is to have approximately the same chance of recording the score sequence of one of the  $k$ -fold secondary trials we may argue as follows. Absolute equiprobability implies that the chance of meeting any such

sequence is  $2^{-k}$ ; but we cannot hope to achieve this. All we can hope to achieve is that:

(a) it will lie in a range  $\frac{1 \pm \epsilon}{2^k}$ ;

(b) the maximum proportionate bias  $\epsilon$  will not exceed an agreed limit, e.g. 0.01 (1 per cent).

Now the probability of scoring 1 or 0 at a single trial lies in the range  $\frac{1}{2}(1 \pm b^r)$ . So the worst that can happen is that the probability of getting a particular  $k$ -fold sequence lies in the range:

$$\left(\frac{1 \pm b^r}{2}\right)^k \simeq \frac{1 \pm kb^r}{2^k}$$

Thus the condition we seek is that  $(1 + \epsilon) \simeq (1 + kb^r)$  or  $\epsilon \simeq kb^r$ .

Illustratively, we may suppose that

(a) the size of the group is  $N = 73$  so that  $2^6 < N \leq 2^7$  and  $k = 7$ ;

(b) we are confident that  $p$  lies between 0.6 and 0.4, in the sense that the maximum single toss bias allowable is 20 per cent, i.e.  $b = 0.2$ ;

(c) we are satisfied if the maximum proportionate bias of the  $k$ -fold sequence is 1 per cent, i.e.  $\epsilon = 0.01$ .

In that event we shall choose  $r$  so that  $0.01 = 7 (0.2)^r$ . Since  $7 (0.2)^3 = 0.016$  and  $7 (0.2)^4 = 0.00224$ , it will ensure that  $\epsilon \leq 0.01$ , if  $r = 4$ , and the allocation of an individual will involve  $4(7) = 28$  tosses of the coin.

2. *It is improper to speak of the probability that a verdict on the outcome of a single trial is true. We can speak with propriety only of the frequency of correct assertion in an unending sequence of trials, and then only if we adhere consistently to the same rule. Since the theory of probability conceived in the foregoing terms does not sanction the right to vary the rules of the game in accordance with the player's luck, we then shoulder the obligation to state in advance the rule in its entirety.*

The second thesis I have proposed carries with it a corollary that has only very lately gained recognition, and by no means universally. In so far as we can claim any usefulness for a rule of the type under discussion, the statement of it must include a specification of the size of the sample. Now the self-same property which bestows on such sample distributions as the  $t$  and the  $F$  (or Fisher's  $z$ ) a peculiar fascination from the viewpoint of the mathematician deprives us of the opportunity of doing so. If we accept my second thesis, we have thus to admit that we have as yet no means at our disposal for assigning a meaningful confidence interval: (a) to a Gaussian point-estimate in its proper domain of the calculus of error; (b) to a mean or other parameter of a table set out in accordance with the Fisher prescription for *Analysis of Variance*. By the same token, the test battery of the technique designated *Analysis of Covariance* forfeits any claims to usefulness endorsed by a calculus of judgments consistent with the Forward Look.

In short, the attempt of Neyman and of Wald to rehabilitate the theory of test procedure and of interval estimation in terms consistent with what Neyman himself calls inductive behaviour has led to an admitted (p. 438) *impasse*. It forced Wald to enunciate the *Minimax* solution which is wholly arbitrary, embodying a concept of decision ostensibly behaviouristic but on closer examination consistent only with the view which locates probability in the mind. I do not venture to express any opinion about whether mathematical ingenuity will eventually surmount the difficulty Neyman himself concedes; but it seems to me that we have now before us a simple choice. If we decline to relinquish the mathematical apparatus of current test and interval estimation, we must retreat from a behaviourist approach to the credentials of statistical theory. In that event, we shall embrace arbitrary axioms the truth of which we can never hope to prove conclusively. If we decline a profession of faith without justification by works, we must abandon the mathematical apparatus of sampling distributions now in vogue. In that event we must dig more secure foundations on which to build a new edifice of theory.

The choice here stated raises issues of which the implications extend far beyond the confines of current statistical con-

troversy. We have seen (p. 443) that one writer prefers to announce a new law of thought rather than to abandon Fisher's concept of a fiducial distribution. Among those who embrace the new law of thought no single exponent has come forward to vindicate its claim to consideration for any reason other than the plenary absolution it confers on adherence to a calculus conceived in solipsistic terms. Thus the dilemma forced on us by the attempt to accommodate the algebraic convenience of the  $t$ -distribution with the logical requirements of a concept of confidence consistent with an empiricist viewpoint has brought into sharper relief an as yet unresolved antinomy between the traditional outlook of the experimental scientist and the current metaphysic of the professional statistician. The antinomy itself is not new. It was latent in the entire tradition of thought which Quetelet, Galton and Pearson bequeathed to our generation. In his own day, Claude Bernard recognised it as such.

3. *Any proper terms of reference we may claim for a calculus of error, for a calculus of exploration or for a calculus of judgments restricts their legitimate use to situations of which we can predicate a fixed framework of repetition.*

That such a postulate is admissible in the domain of error *sensu stricto* is not open to question; but the family likeness of the Gaussian algebra and that of multivariate or of factor analysis tempts us to forget that a common algebraic symbolism furnishes no safeguard for conserving a necessarily implicit, if not actually explicit, assumption of the stochastic calculus of error. In the new role which Quetelet, Galton and Pearson found for the Gaussian algebra, we have to relegate the randomising procedure to Nature; and we have to abandon the opportunity of recording the same observation in precisely the same circumstances on successive occasions. For the classical theory of wagers we can define randomness in a meaningful way only if we also predicate a temporal framework of limitless repetition. This may be admissible in the domain of populations of living beings when we are dealing with highly inbred stocks in an ideally standardized environment, though then only if we neglect mutation as a second order effect. Otherwise,

our opportunities of observation refer only to events which are historically unique.

The *raison d'être* of the infinite hypothetical population bequeathed by Laplace to Quetelet's successors including the contemporary school of R. A. Fisher, is the search for a formula to sidestep this dilemma. The undertaking fails at two levels. One is that the conceptually static and infinite population endorses the gratuitous belief that blindfold selection is *ipso facto* randomwise; but the construction or use of tables of random numbers by those who subscribe to it sufficiently shows that they entertain grave doubts about it. This is confessedly an argument *ad hominem*; but the concept of a static infinite population is also exceptionable at an operational level. Any conclusions to which statistical procedures may lead us are useful only if we can legitimately identify situations in which we propose to apply them with the entire class of situations subsumed by the putative fixed framework of repetition. Needless to say, it is commonly, if not invariably, impossible to conceive of an infinite actual population which we can pin down to such definite requirements.

\*                      \*                      \*                      \*

The foregoing theses epitomise the positive outcome of the task undertaken in this book; but they do not impinge on the issue: what light can statistical studies based as such on populations shed on the mechanism of *individual* behaviour? If we abandon the claim to assign a probability to the truth of a particular verdict as inconsistent with a behaviourist approach, we must regard a calculus of judgments as a prescription for a collective discipline of interpretation in contradistinction to a sanction for individual conviction. We may still concede the possibility of building an edifice of new theory on foundations more secure than the infinite hypothetical population; but we have then to ask: what useful role can such a calculus fulfil in the biological or social sciences? For my part, I should hesitate to say *none at all*; but we shall be wise to refrain from indulging in unjustifiable expectations about the situations in which it may be helpful to enquiry. We

shall be still wiser, if our revaluation of the claims of a calculus of judgments encompasses a re-examination of the type of situations to which it has no relevance.

To do justice to the foregoing query, we must first ask ourselves: to what class of questions do we seek an answer when we invoke such procedures as statistical tests and interval estimation? A foregoing citation (p. 227) from Claude Bernard in the context of the clinical trial forewarns us of and forearms us against one widely current delusion about their relevance. If it is proper to say that statistics is the science of averages, we should be hesitant about enlisting statistics in experimental enquiry before asking: is the answer to our question usefully expressible in terms of an average? Otherwise, undue reliance on statistics must impede the march of science by encouraging us to ask useless questions or to refrain from asking the most useful ones.

Fisher's oft-quoted parable about the lady and the tea-cup is an instructive text for this theme, since it illustrates an issue of fundamental importance to the study of animal or human behaviour and one with topical relevance *vis à vis* latter-day claims concerning the existence of so-called *extrasensory perception*. As R. A. Fisher states it, the problem is as follows. A lady declares that she can distinguish by taste whether her hostess puts milk in the cup before the tea or *vice versa*. He proposes to test the truth of her assertion by allowing her to taste 8 cups after telling her that he has mixed four in one way and four in the other. From the writer's viewpoint, this form of words is exceptionable, if we reject the relevance of stochastic induction to the outcome of single trials; but this need not here detain us. The possible answers the lady may give, when asked to classify the 8 cups on the assumption stated, subsume  $(8! \div 4! \ 4!) = 70$  permutations, only one of which can correspond to actuality. If we deem denial of her claim to discriminate as equivalent to correct assessment with a long-run frequency consistent with truly random choice, the denial of her claim (Fisher's null hypothesis) will occur with a frequency of one in seventy. If she discriminates consistently, she will always be right. In effect, therefore, Fisher proposes the rule: say that she can discriminate if the outcome is correct

assignment. Then  $70^{-1}$  is the probability of error, if she cannot discriminate at all, and zero is the probability of error, if she can. Thus  $P_f \leq 70^{-1}$ .

If the only admissible alternatives are as stated, the positive outcome of several such experiments without a failure will presumably satisfy the sceptic to whom the investigator submits his findings, and he will not need to lean on the statistician for support; but Fisher himself envisages a lady content to claim "not that she could draw the distinction with invariable certainty, but that, though sometimes mistaken, she would be right more often than not." Evidently, this claim is not susceptible to discussion in statistical terms, unless we add one or both of two emendations: (a) she can herself state her claims with more precision; (b) the jury can agree about what numerical level of long-run success is worthy of consideration.

Neyman makes this point vigorously in a critique of Fisher's treatment of the problem. For him, as for Fisher, the probability of correctly identifying a single cup is  $p = \frac{1}{2} = p_0$ , if she has no discrimination at all; but we can assign the appropriate sample size to endorse an acceptable uncertainty safeguard for a rule of procedure only if we assign a positive number  $k < \frac{1}{2}$ , being content to deny that she can discriminate unless  $p \geq p_0 + k = p_k$ . In short, the lady then supposedly behaves like a biased penny, and we can formulate a decision rule to test its bias ( $k$ ) only if we first decide to dismiss certain values of  $k$  as trivial.

If we adopt a somewhat naïve view about a biased penny, this disposes of the dilemma; but any such conceptualisation of a *degree of sensory perception* forces Neymann to postulate "identity of conditions and the complete independence of the  $n$  successive classifications of pairs of cups of tea." Some of my readers will therefore share Wrighton's misgivings, when he asks: "If conditions were in fact identical, what factor would intervene to make the lady's response vary?" Wrighton himself answers it as follows:

The only formal explanation is that in this respect the lady's sensory equipment operates in a fashion characteristic of the roulette board. Alternatively, following Fisher's treatment of other problems,



we may identify *the lady as she is at the time of the experiment* with a conceptual lady out of whose infinitely protracted tea-tasting the experience of the experiment is regarded as a random sample. The idea may be attractive, but it carries with it an embarrassing consequence (if we pursue Neyman's illustration). If the experiment demonstrates the phenomenon, it is the conceptual lady who must in fairness be rewarded, and if not, it is the conceptual lady whose pretensions must be exposed.

For Neyman, however, the  $p$  appears to be real. The successive trials of the lady are specifically postulated as independent and she is axiomatically endowed with an unvarying  $p$ . It is difficult to imagine a justification which would obtain general acceptance among biologists and experimental psychologists for either of these postulates.

Some may feel that the remarks last cited do less than justice to Neyman's treatment of the problem. In my view, we shall do justice to it only if we recognise a confusion of aims, when we enlist a decision rule procedure in this context. To me as a biologist, Neyman's procedure is intelligible only if the end in view is *personnel selection*, in this context picking out ladies who can invariably discriminate from a group containing others who cannot. We thus come back to the issue: to what class of questions do we seek an answer when we invoke statistical procedures?

When discussing populations our concern may be: (a) to classify the constituent individuals in a rough and ready way for convenience of administrative book-keeping; (b) to arrive at a deeper understanding of the laws of individual behaviour, i.e. a confident prescription for evoking particular responses. If the end in view is a piece of social accountancy, the answer we shall seek will be an average or will be justifiable in terms of a criterion of average success. An average level of performance may then be consonant with the operational intention. If the end in view is to arrive at a useful understanding of individual behaviour, what peculiarities of nature and previous nurture determine whether different individuals respond to the same stimulus in different ways, or what internal states and external circumstances determine the different responses of one and the same individual to one and the same stimulus are the questions to which we rightly seek an answer. We

have indeed accomplished our objective only if we can say of such and such an individual that the probability of responding in such and such a way and in such and such circumstances does not appreciably differ from unity or from zero. In the context of research on the mechanism of behaviour, Neyman's  $p$  lying somewhere between 0.5 and 1.0 is a tiresome distraction which contributes nothing to the business in hand.

Such is the glamour of statistics, that the oncoming generation too readily overlook this distinction. Procedures which might be defensible as screening devices thus become an encouragement to evade reality and an excuse for curbing curiosity about fundamental issues. In short, a calculus of uncertainty is becoming the creed of a cult which disdains to press forward to greater certainty when certainty is indeed attainable, and when nothing short of certainty constitutes a useful addition to the enduring corpus of scientific knowledge. In part, the appeal of its doctrine is due to the prestige which mathematics, however irrelevant, confers on those who use it as a tool of interpretation. In part, it results from a failure to distinguish between the permanent factual bricks continually added to the ever-growing edifice of scientific knowledge and a temporary scaffolding of metaphors which the builders discard at their own convenience.

What most helps to perpetuate the confusion last stated is a monistic prejudice which demands a unitary formula for scientific method. Having found no such formula, I have resigned myself to an eclectic acceptance of a diversity of scientific methods; and I believe that it is high time to unmask the semantic implications of using terms such as law, theory, hypothesis in the diverse domains of the retrospective and prospective disciplines. That it leads to confusion of thought, the reader will readily enough concede if he or she reflects on a few familiar examples of their use. Thus the *Cell theory* in biology subsumes merely the factual content of our observational record; but the *Wave theory* of light subsumes a metaphor congenial to the exposition of the metrical relations we encounter in the study of optical interference and the phenomenon of double refraction. Mendel's *law* of segregation is a recipe for action in the practice of plant breeding and

runs the gauntlet of possible disproof whenever we apply it. The *law* of evolution is an obituary notice, interpreting *past* change in the light of *current* experience. As such, it offers us no opportunity to test its validity in the domain of action.

One might multiply such examples indefinitely, and if we draw what seems to me to be the inescapable conclusion, it will be very difficult to see how a stochastic calculus of exploration consistent with the Forward Look can claim an enduring place among scientific methods of proven usefulness. Briefly, then, the outcome of our examination of the present crisis in statistical theory from a behaviourist viewpoint which encompasses no axioms unwarranted by the practical experience of mankind is as follows:

(a) the credentials of a stochastic calculus of aggregates stand secure from the impact of any contemporary debatable issues arising from disagreement concerning the proper terms of reference of a theory of probability;

(b) in its legitimate and original domain of error, the procedure of point estimation is at least a meaningful convention;

(c) the factual basis for the assumptions inherent in such exploratory and descriptive procedures as multivariate analysis and factor analysis is exiguous and the ostensible rationale is difficult if not impossible to reconcile with any definable historic framework of randomwise repetition;

(d) the available theory of sampling distributions does not suffice to endorse a non-subjective theory of test procedure or of interval estimation in situations which admit of recourse to randomising devices;

(e) if it proves possible to rehabilitate a stochastic calculus of judgments in terms consistent with the full implications of a behaviourist outlook, it will be the more necessary to remind ourselves that it can merely endorse assertions which encourage us to probe more deeply, and without reliance thereon, into nature's secrets;

(f) the claim of the statistician to prescribe the design of experiments in accordance with the requirements of significance test procedures and of fiducial estimates is con-

sistent neither with a behaviourist view of the proper scope of stochastic induction nor with what should presumably be the primary intentions of the investigator.

There will assuredly be few to whom so iconoclastic a verdict will be palatable; and the writer himself has reached it reluctantly with all the mental discomfort of discarding a weighty incrustation of prevalent custom-thought. In following to what now seems to me to be the bitter end of a trail which others, more especially J. Neyman, E. S. Pearson and A. Wald, have blazed, I have at all times hoped that the prospect of the promised land would prove to be more fertile. It is not agreeable to cherish opinions which isolate one from the bulk of one's intellectual contemporaries; but if one is to enjoy the privileges of intellectual adventure, one must resign oneself to a little isolation as the price of self-indulgence in so exotic a luxury as intellectual rectitude. If I am in error, I can at least hope that the issues I have brought before my readers will compel others wiser than myself to formulate more clearly than heretofore a rational basis for convictions which my own reflections have forced me to abandon. Statistical theory will then enlarge its claim to our respect by relinquishing authoritarian pretensions uncongenial to the temper of science. Contrariwise, the possibility that I may be right in the main does not discourage me. Only when their elders have cleared the site for new foundations will a younger generation be intellectually free to undertake truly creative work now neglected and denigrated.

Persuaded of this conviction, I conclude by citing a challenge for rebuttal or otherwise. In his *Cours de Philosophie Positive*, Auguste Comte (1839), who led the contemporary revolt against Quetelet's *physique sociale*, expresses himself in the following terms:

La seule aberration de ce genre qui eût pu mériter quelque discussion sérieuse, si l'ensemble de ce Traité ne nous en avait d'avance radicalement dispensé, c'est la vaine prétention d'un grand nombre de géomètres à rendre positives les études sociales d'après une subordination chimérique à l'illusoire théorie mathématique des chances. C'est là l'illusion propre des géomètres en philosophie politique, comme celle des biologistes y consiste surtout,

ainsi que je l'ai ci-dessus expliquée, à vouloir ériger la sociologie en simple corollaire ou appendice de la biologie, en y supprimant, dans l'un et l'autre cas, l'indispensable prépondérance de l'analyse historique. Il faut néanmoins convenir que l'aberration des géomètres est, à tous égards, infiniment plus vicieuse et beaucoup plus nuisible que l'autre; outre que les erreurs philosophiques quelconques sont, en général, bien autrement tenaces chez les géomètres, directement affranchis, par la haute abstraction de leurs travaux de toute subordination rigoureuse à l'étude réelle de la nature. . . . Serait-il possible, en effet, d'imaginer une conception plus radicalement irrationnelle que celle qui consiste à donner pour base philosophique, ou pour principal moyen d'élaboration finale, à l'ensemble de la science sociale, une prétendue théorie mathématique, où, prenant habituellement des signes pour des idées, suivant le caractère usuel des spéculations purement métaphysiques, on s'efforce d'assujétir au calcul la notion nécessairement sophistiquée de la probabilité numérique, qui conduit directement à donner notre propre ignorance réelle pour la mesure naturelle du degré de vraisemblance de nos diverses opinions? Aussi aucun homme sensé n'a-t-il été, dans la pratique sociale, effectivement converti de nos jours à cette étrange aberration, quoique sans pouvoir en démêler le sophisme fondamental. Tandis que les vraies théories mathématiques ont fait, depuis un siècle, de si grands et si utiles progrès, cette absurde doctrine, sauf les occasions de calcul abstrait qu'elle a pu susciter, n'a véritablement subi, pendant le même temps, malgré de nombreux et importants essais, aucune amélioration essentielle, et se retrouve aujourd'hui placée dans le même cercle d'erreurs primitives, quoique la fécondité des conceptions constitue certainement, à l'égard d'une science quelconque, le symptôme le moins équivoque de la réalité des spéculations.

If Comte is right in this declaration, what does the future hold for the practitioner of statistics as an academic discipline in the domain of the biological and of the social sciences? The question admits of no simple and, to most of my contemporaries, of no agreeable answer. The stochastic theory of genetical populations remains an outstanding achievement of experimental biology, and time may conceivably come when a calculus of aggregates can usefully embrace other cellular phenomena. As yet we can discern no prospect of this in the social sciences. Unless it is possible to rebuild a calculus of judgments on a firmer basis it may well be that the future

of statistics in the social sciences is what Anscombe (p. 14) contemptuously calls statistics as the term is used by some continental demographers.

If so, it may still claim a modest field of usefulness by prescribing techniques of summarisation appropriate to different types of problems which arise in the study of populations. The first volume of Yule's treatise claimed to do little more. Stochastic models may still suggest recipes for summarising indices which do indeed summarise as do the rank coefficients of Spearman and of Kendall; but a more parsimonious view of the relevance of the calculus of probability to social occurrence is not likely to offer attractive opportunities for the exercise of mathematical subtlety of a high order. The focus of contemporary controversy has been the nature of statistical inference subsumed by what I have called a calculus of judgments; and the vehemence of the debate has demonstrated its vulnerability. What has emerged from these pages is that the stochastic theory of curve fitting subsumed by what I have called a calculus of exploration is even more vulnerable, if we scrutinise its credentials on the same footing. That graduating devices have a limited usefulness in the study of populations as first aids in the search for deeper insight we may well concede; but those who adhere to a behaviourist view will be willing to endorse a stochastic rationale for their prescription only if it is possible to rehabilitate their credentials without recourse to such Platonic constructs as the infinite hypothetical population, the normal man and the normal environment.

# APPENDIX ONE

## CENTRAL LIMIT THEOREM

BY DEFINITION, we speak of a score  $x$  whose mean value is  $M$  as a normal variate, if the equation of its frequency ( $y$ ) in the interval  $x \pm \frac{1}{2}dx$  is

$$y_x = \frac{1}{\sqrt{2\pi V}} \exp \left[ -\frac{(x - M)^2}{2V} \right] \cdot dx$$

In this expression the constant  $V$  is the variance (*second mean moment*) of the distribution. If we denote by  $X = (x - M)$  the deviation of  $x$  from its mean value, there will correspond to each value of  $x$  with the same frequency in the appropriate interval a *standard score*  $(X \div \sqrt{V}) = c$ . The equation for the frequency of  $c$  in the interval  $c \pm \frac{1}{2}dc$  is

$$y_c = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2}c^2 \right) \cdot dc$$

For any *discrete* distribution of a *standard score*  $c = (X \div \sqrt{V})$  we define the  $k$ th mean moment in terms of  $f_c$  the frequency of  $c$  in the interval  $c \pm \frac{1}{2}dc$  as the mean value of the  $k$ th power of  $c$ , viz.:

$$M_k = \sum_{-\infty}^{+\infty} f_c \cdot c^k = E(c^k) \quad . \quad . \quad . \quad (i)$$

Whence, alternatively:

$$M_k = V^{-\frac{1}{2}k} \cdot E(X^k)$$

If  $f_c = y_c \cdot dc$  for a continuous variate, we write

$$M_k = \int_{-\infty}^{+\infty} y_c \cdot c^k \cdot dc$$

For a normal variate

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}c^2} \cdot c^k \cdot dc \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}c^2} \cdot c^k \cdot dc \quad (\text{if } k \text{ is even}) \text{ or zero (if } k \text{ is odd)} \end{aligned}$$

To evaluate this integral, we may substitute  $c^2 = Q$ , so that  $dc = \frac{1}{2}Q^{-\frac{1}{2}}.dQ$ , and

$$M_k = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{1}{2}Q} \cdot Q^{\frac{1}{2}(k-1)} \cdot dQ$$

This is a Gamma function whose value is zero for *odd* values of  $k$ , and its value for even values of  $k$  is the product  $1 \cdot 3 \cdot 5 \dots (k-1)$ . For the normal distribution we therefore have:

$$\left. \begin{array}{l} M_1 = 0 ; M_3 = 0 ; M_5 = 0 ; M_7 = 0 \\ M_2 = 1 ; M_4 = 3 ; M_6 = 15 ; M_8 = 105 \end{array} \right\} \quad . \quad (ii)$$

Let us now consider a discrete distribution of unit trial scores ( $x$ ) about mean  $M$  with variance  $V = M_2(1)$  which is by definition the mean value of  $(x - M)^2 = X^2$ . We shall assume that unit trials are independent. For the  $r$ -fold sample score sum we shall write  $x_r$  and the corresponding deviation  $X_r = (x_r - rM)$  whose mean value is zero. For the  $(r+1)$ th unit trial score we shall write  $x_1$  and for the  $(r+1)$ -fold sample score sum  $x_{r+1}$ , so that  $X_{r+1} = (x_r + x_1) - (rM + M) = X_r + X_1$ . The  $k$ th mean moment of the  $(r+1)$ -fold sample score sum is then by definition

$$\begin{aligned} m_k(r+1) &= E(X_r + X_1)^k \\ &= \sum_{p=0}^{p=k} E(k_{(p)} \cdot X_r^{k-p} \cdot X_1^p) \end{aligned}$$

In virtue of independence of the unit trial scores, therefore

$$\begin{aligned} m_k(r+1) &= \sum_{p=0}^{p=k} k_{(p)} \cdot E(X_r^{k-p}) \cdot E(X_1^p) \\ \therefore m_k(r+1) &= \sum_{p=0}^{p=k} k_{(p)} \cdot m_{k-p}(r) \cdot m_p(1) \end{aligned}$$

When  $p = 1$ ,  $m_p(1) = 0$  and when  $(k-p) = 1$ ,  $m_{k-p}(r) = 0$ . Whence we have

$$\begin{aligned} m_2(r+1) &= m_2(r) + m_2(1) \\ m_3(r+1) &= m_3(r) + m_3(1) \\ m_4(r+1) &= m_4(r) + 6m_2(r) \cdot m_2(1) + m_4(1) \\ &\text{etc. etc.} \end{aligned}$$



# APPENDIX I

By successively putting  $r = 1, 2, 3$ , etc., in the above we derive:

$$m_2(2) = 2m_2(1) \quad ; \quad m_2(3) = 3m_2(1) \quad \text{etc.}$$

$$m_3(2) = 2m_3(1) \quad ; \quad m_3(3) = 3m_3(1) \quad \text{etc.}$$

$$m_4(2) = 2m_4(1) + 6m_2^2(1) \quad ; \quad m_4(3) = 3m_4(1) + 18m_2^2(1) \quad \text{etc.}$$

$$m_4(4) = 4m_4(1) + 36m_2^2(1) \quad . . . \quad \text{etc.}$$

Thus more generally:

$$m_2(r) = r.m_2(1) \quad ; \quad m_3(r) = r.m_3(1)$$

$$m_4(r) = r.m_4(1) + 3r^{(2)}.m_2^2(1)$$

By definition,  $m_2(1) = V$  in the above is the variance of the unit sample distribution and  $m_2(r) = V_r$  that of the  $r$ -fold score sum. The  $k$ th mean moment of the  $r$ -fold score sum distribution in standard form is by definition

$$M_k(r) = V_r^{-\frac{1}{2}k}.m_k(r)$$

$$\therefore M_2(r) = 1 \quad ; \quad M_3(r) = \frac{m_3(1)}{\sqrt{rV^3}}$$

$$M_4(r) = \frac{m_4(1)}{rV^2} + \frac{3(r-1)}{r}$$

For large values of  $r$ , we therefore have

$$M_2(r) = 1 \quad ; \quad M_3(r) \simeq 0 \quad ; \quad M_4(r) \simeq 3$$

In the same way, we can show that for large values of  $r$  higher moments approach the corresponding values for the normal distribution:

$$M_5(r) \simeq 0 \simeq M_7(r) \quad ; \quad M_6(r) \simeq 15 \quad ; \quad M_8(r) \simeq 105 \quad \text{etc.}$$

## APPENDIX TWO

### THE UMPIRE BONUS MODEL

CERTAIN RELATIONS between a system of paired scores  $x_a, x_b$  are tautologies which do not depend on stochastic considerations. They are easily deducible by recourse to the operators

$E_{a.b}(f_{a.b})$  arithmetic mean of any function  $f_a$ , of  $x_a$  for a fixed value of  $x_b$ .

$E_{b.a}(f_{b.a})$  ditto  $f_b$  of  $x_b$  for a fixed value of  $x_a$ .

$E_a(f_{ab})$  arithmetic mean of any function  $f_{ab}$  of both  $x_a$  and  $x_b$  for all values of  $x_a$ .

$E_b(f_{ab})$  ditto for all values of  $x_b$ .

$E(f_{ab})$  ditto for all values of both  $x_a$  and  $x_b$ .

From the definition of the arithmetic mean :

$$E_a.E_{b.a}(f_{ab}) = E(f_{ab}) = E_b.E_{a.b}(f_{ab})$$

$$E(f_a) = E_a(f_a) \quad ; \quad E(f_b) = E_b(f_b)$$

The following definitions of summarising indices in this notation are necessary for what follows.

*Means:*

$$E_a(x_a) = M_a \quad ; \quad E_{a.b}(x_a) = M_{a.b} \quad ; \quad E_b(x_b) = M_b \quad ; \quad E_{b.a}(x_b) = M_{b.a}$$

$$E_a(M_{b.a}) = M_b \quad \text{and} \quad E_b(M_{a.b}) = M_a$$

*Variances:*

$$E_a(x_a - M_a)^2 = V_a = E_a(x_a^2) - 2M_a.E_a(x_a) + M_a^2 = E_a(x_a^2) - M_a^2$$

Similarly :

$$E_{a.b}(x_a - M_{a.b})^2 = V_{a.b} = E_{a.b}(x_a^2) - M_{a.b}^2$$

$$E_{b.a}(x_b - M_{b.a})^2 = V_{b.a} = E_{b.a}(x_b^2) - M_{b.a}^2$$

$$E_b(x_b - M_b)^2 = V_b = E_b(x_b^2) - M_b^2$$

For brevity, it is convenient to write:

$$X_a = x_a - M_a ; X_{a.b} = x_{a.b} - M_{a.b} \quad \text{etc.}$$

If  $x_k = kx_a$ , so that  $M_k = kM_a$ , it follows that

$$V_k = E(x_k - M_k)^2 = k^2 V_a$$

*Covariance:*

$$\text{Cov } (x_a, x_b) = E(X_a \cdot X_b) = E(x_a \cdot x_b) - M_a \cdot M_b$$

$$E(x_a \cdot x_b) = E_a \cdot E_{b.a}(x_a \cdot x_b) = E_a(x_a \cdot M_{b.a})$$

$$\therefore E_a(x_a \cdot M_{b.a}) = \text{Cov } (x_a, x_b) + M_a \cdot M_b = E_b(x_b \cdot M_{a.b})$$

*Linear Regression of  $x_a$  on  $x_b$*

$$M_{a.b} = k_{ab} \cdot x_b + C \quad . \quad . \quad . \quad . \quad . \quad (i)$$

$$M_a = E_b(M_{a.b}) = k_{ab} \cdot E_b(x_b) + C = k_{ab} \cdot M_b + C$$

$$\therefore (M_{a.b} - M_a) = k_{ab} \cdot X_b \quad . \quad . \quad . \quad . \quad . \quad (ii)$$

$$\therefore E_b(X_b \cdot M_{a.b}) - M_a \cdot E_b(X_b) = k_{ab} \cdot E(X_b^2)$$

Since  $E_b(X_b) = 0$  by definition

$$k_{ab} \cdot V_b = \text{Cov } (x_a, x_b) \quad \text{and} \quad k_{ab} = \frac{\text{Cov } (x_a, x_b)}{V_b} \quad (iii)$$

*Linear Regression of  $x_b$  on  $x_a$*

$$M_{b.a} - M_b = k_{ba} \cdot X_a$$

$$k_{ba} \cdot V_a = \text{Cov } (x_a, x_b) \quad \text{and} \quad k_{ba} = \frac{\text{Cov } (x_a, x_b)}{V_a} \quad (iv)$$

Whence, if  $V_a = V_b$ ,  $k_{ba} = k_{ab}$ .

*Product-Moment Correlation Coefficient.*

$$r_{ab} = \frac{\text{Cov } (x_a, x_b)}{\sqrt{V_a \cdot V_b}} \quad . \quad . \quad . \quad . \quad . \quad (v)$$

When regression is linear in both dimensions:

$$r_{ab}^2 = k_{ab} \cdot k_{ba} \quad . \quad . \quad . \quad . \quad . \quad (vi)$$

If  $k_{ba} = k_{ab}$ ,  $k_{ab} = r_{ab} = k_{ba}$

. . . . .

*The Umpire Bonus Model:* We now postulate a system of scores  $x_u$ ,  $x_{a.0}$  and  $x_{b.0}$  which are independent stochastic variables, so that

$$\text{Cov } (x_{a.0}, x_u) = \text{Cov } (x_{b.0}, x_u) = 0 = \text{Cov } (x_{a.0}, x_{b.0})$$

We then define

$$x_a = A \cdot x_u + x_{a.0} \quad \text{and} \quad x_b = B \cdot x_u + x_{b.0} \quad . \quad (\text{vii})$$

It follows that:

$$M_a = A \cdot M_u + M_{a.0} \quad \text{and} \quad M_b = B \cdot M_u + M_{b.0}$$

$$\therefore X_a = A \cdot X_u + X_{a.0} \quad \text{and} \quad X_b = B \cdot X_u + X_{b.0}$$

$$E(X_a \cdot X_u) = A \cdot E(X_u^2) + E(X_u \cdot X_{a.0})$$

$$E(X_b \cdot X_u) = B \cdot E(X_u^2) + E(X_u \cdot X_{b.0})$$

$$E(X_a \cdot X_b) = AB \cdot E(X_u^2) + A \cdot E(X_u \cdot X_{b.0}) \\ + B \cdot E(X_u \cdot X_{a.0}) + E(X_{a.0} \cdot X_{b.0})$$

Whence we obtain:

$$\text{Cov } (x_a, x_u) = A \cdot V_u ; \quad \text{Cov } (x_b, x_u) = B \cdot V_u ;$$

$$\text{Cov } (x_a, x_b) = AB \cdot V_u$$

$$\therefore r_{au} = \frac{A \cdot V_u}{\sqrt{V_a \cdot V_u}} = A \sqrt{V_u \div V_a} ;$$

$$r_{bu} = \frac{B \cdot V_u}{\sqrt{V_b \cdot V_u}} = B \sqrt{V_u \div V_b} ; \quad r_{ab} = \frac{AB \cdot V_u}{\sqrt{V_a \cdot V_b}} \quad (\text{ix})$$

$$\therefore r_{ab} = r_{au} \cdot r_{bu} \quad . \quad . \quad . \quad . \quad . \quad (\text{x})$$

If  $A = 1 = B$ ,  $r_{ab} = 0.5$  when  $V_a = V_b$  and  $V_{a.0} = V_u$ ; and if regression is also linear in both dimensions  $k_{ab} = 0.5 = k_{ba}$ .

*Statistical Independence.* The mean of the score sum  $x_s = x_a + x_b$  is  $M_s = M_a + M_b$  and its variance is by definition

$$\begin{aligned} E(x_s - M_s)^2 &= E(x_a^2 + x_b^2 + 2x_a \cdot x_b - 2x_a \cdot M_b - 2x_b \cdot M_a \\ &\quad - 2x_a M_a - 2x_b M_b + M_a^2 + M_b^2 + 2M_a M_b) \\ &= E(x_a^2) - M_a^2 + E(x_b^2) - M_b^2 + 2E(x_a \cdot x_b) - 2M_a M_b \\ &= V_a + V_b + 2 \text{Cov}(x_a, x_b) \end{aligned}$$

When  $x_a$  and  $x_b$  are independent stochastic variates:

$$\begin{aligned} E(x_a \cdot x_b) &= E(x_a) \cdot E(x_b) = M_a \cdot M_b \\ \therefore \text{Cov}(x_a, x_b) &= 0 \end{aligned}$$

Whence for two independent variates  $x_a, x_b$  the variance ( $V_s$ ) of the score sum is:

$$V_s = V_a + V_b \quad . \quad . \quad . \quad . \quad . \quad (xi)$$

From (xi) we see that

$$V_a = A^2 V_u + V_{a.0} ; \quad V_b = B^2 V_u + V_{b.0}$$

Accordingly, we may write:

$$\begin{aligned} \frac{A^2 V_u}{V_a} + \frac{V_{a.0}}{V_a} &= 1 = r_{au}^2 + \frac{V_{a.0}}{V_a} \\ \frac{B^2 V_u}{V_b} + \frac{V_{b.0}}{V_b} &= 1 = r_{bu}^2 + \frac{V_{b.0}}{V_b} \end{aligned}$$

We may therefore speak of  $r_{au}^2$  and  $r_{bu}^2$  respectively as the contribution of the umpire bonus to the total variance of the score of player A and of player B.

If either A or B is zero or if both are zero  $x_a = x_{a.0}$ , the numerator of  $r_{ab}$  in (ix) is zero, whence  $r_{ab} = 0$ . In that event  $x_a$  and  $x_b$  are independent. For any two independent variates  $x_a$  and  $x_b$ , the product-moment coefficient  $r_{ab}$  must be zero, since in that event also  $\text{Cov}(x_a, x_b) = 0$ . If  $x_{a.0} = 0 = x_{b.0}$ , so that  $B \cdot x_a = A \cdot x_b$  there is perfect linear correspondence between the scores of the players, and

$$V_a = A^2 V_u ; \quad V_b = B^2 V_u ; \quad \sqrt{V_a \cdot V_b} = AB \cdot V_u$$

Thus  $r_{ab} = +1$  if the signs of A and B are alike and  $r_{ab} = -1$  if the signs are unlike. These limits necessarily hold good if  $x_a$  is a perfect linear function of  $x_b$ , i.e.  $x_a = kx_b$  in which event  $Cov(x_a, x_b) = E(kb^2) - kM_b^2 = kV_b$  and  $V_a = k^2V_b$ , so that  $\sqrt{V_a \cdot V_b} = kV_b$ .

## APPENDIX THREE

### DETERMINANT NOTATION

WE DEFINE A DETERMINANT of order  $n$  as a set of  $n^2$  numbers laid out gridwise in  $n$  rows and  $n$  columns; and define the numerical value of the determinant of order 2 as

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1$$

The definition of the numerical value of any determinant of order higher than 2 depends on the application of a rule for expressing a determinant of order  $n$  in terms of determinants of order  $(n - 1)$  and hence by iteration in terms of determinants of order 2. Corresponding to each of the  $n$  elements in the top row, we speak of the *minor* of order  $(n - 1)$  as the residual determinant formed by eliminating the elements of the top row and of the corresponding column. The rule for reduction is as follows: proceeding from left to right multiply each minor by the corresponding element in the top row assigning positive and negative values to the products alternately. To represent the rule compactly it is convenient to label a determinant by its diagonal terms thus

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = \Delta(a_1 \cdot b_2 \cdot c_3 \cdot d_4)$$

The rule for reduction is then:

$$a_1 \cdot \Delta(b_2 c_3 d_4) - b_1 \cdot \Delta(a_2 c_3 d_4) + c_1 \cdot \Delta(a_2 b_3 d_4) - d_1 \cdot \Delta(a_2 b_3 c_4)$$

We expand each of the minors of order 3 in the same way, e.g.

$$\begin{vmatrix} b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} = b_2 \cdot \Delta(c_3 d_4) - c_2 \cdot \Delta(b_3 d_4) + d_2 \cdot \Delta(b_3 c_4) \\
 = b_2(c_3 d_4 - d_3 c_4) - c_2(b_3 d_4 - d_3 b_4) \\
 + d_2(b_3 c_4 - c_3 b_4)$$

Thus the 4th order determinant reduces to a sum of 12 products involving an element from each of the first two rows and a determinant of order 2. Similarly a 5th order determinant reduces to a sum of 60 products involving an element in each of the first three rows and a determinant of order 2.

In practice, the evaluation of a determinant of high order does not involve the evaluation of many terms. Instead one applies a set of straightforward rules of elimination, subsumed in Aitken's *pivotal condensation* (*vide* Aitken, *Determinants and Matrices*, Oliver and Boyd).



SIGNIFICANCE AS INTERPRETED BY  
THE SCHOOL OF R. A. FISHER

OF THREE WIDELY DIVERGENT VIEWS about the nature of statistical inference, two have hitherto attracted little attention except among professional mathematicians, and have had few protagonists among practical statisticians except—as is true of Neyman's school—in connexion with new recipes for statistical inspection in commerce and manufacture. Contrariwise, the overwhelming majority of research workers in the biological field (including medicine and agriculture), as also a growing body of workers in the social sciences, rely largely on rule of thumb procedures set forth in a succession of manuals modelled on *Statistical Methods for Research Workers* by R. A. Fisher. This publication, which disclaims any attempt to set forth comprehensive derivations to justify the rationale of the author's methods, some of them first announced therein without formal proof, has thus become the bible of a world sect. It has a special importance because the author refers to it in a highly controversial context as a source book of his own *Weltanschauung*.

Some of Fisher's ideas took shape in the context of the early developments of the theory of relativity; and he learnt to move with ease and grace in the empyrean of the hypersphere when abstract multidimensional geometry was still an uncharted domain to theoretical statisticians of an earlier vintage. None of his many self-confessed expositors has attempted the task of interpreting the mathematical techniques involved on a plane intelligible to the investigator with no better equipment than a good practical grasp of the elements of the infinitesimal calculus or to analyse the logical content of the concepts invoked *vis-à-vis* the classical theory of probability. The consumer who is not a trained mathematician has therefore to learn the jargon of *degrees of freedom*, an expression which is meaningful only to those so fortunate as to be familiar with such branches of applied mathematics as the theory of the

gyrostat or the development of thermodynamics in the tradition of Willard Gibbs.

Thus the controversy provoked by the more recent writings of Neyman, Wald, von Mises and others is essentially a challenge to the adequacy of statistical procedures already invoked by most laboratory workers who make much use of statistics; and those who use them most have so far had little opportunity to contribute to a discussion the outcome of which is highly relevant to their day's work. A temperate appraisal of the content of the controversy therefore calls for a more detailed examination of the views of R. A. Fisher and of his disciples than of the more concisely and explicitly stated opinions of his opponents.

To do justice to an evolving system of thought with lexicographical punctiliousness is commonly difficult, if only because it is by no means to the discredit of any philosopher to say that he has changed his views in the course of a prolific career. Happily, we may sidestep this obstacle. Fate has not deprived us of an up-to-date assessment of his mature views, since no sentiment of false modesty has deterred R. A. Fisher from taking the unusual precaution of collecting what he regards as his hitherto major contributions in an impressive quasi-memorial volume with a biographical introduction, with a portrait in photogravure of the author and with his own respectfully retrospective comments on the outstanding importance of each of the literary landmarks re-erected therein.

In these collected *Contributions* (1950) Fisher (35.173a) refers explicitly to the core of the current controversy in the following words:

Pearson and Neyman have laid it down axiomatically that the level of significance of a test must be equated to the frequency of a wrong decision "in repeated samples from the same population." This idea was foreign to the development of tests of significance given by the author in 1925, for the experimenter's experience does not consist in repeated samples from the same population, although in simple cases the numerical values are often the same; and it was, I believe, this coincidence of values in simple cases which misled Pearson and Neyman, who were not very familiar with the ideas of "Student" and the author.

In the references which follow, Fisher quotes under his own name only a short paper (1937) "on a point raised by M. S. Bartlett," his book (1925, 1st edition) *Statistical Methods for Research Workers* and a third as above. Before we consult this source, we may pause to anticipate a curious implication of the foregoing remarks. Our next citation asserts that the static population of the experimenter's experience is also infinite. If so, we exclude the treatment of sampling without replacement from a finite universe from the domain of discourse at the outset, a limitation which is certainly inconsistent with the practice of Fisher's disciples and with the much-quoted tea cup test (*vide infra*) in his later book *Design of Experiments*. Fisher's own analysis of the lady and the tea-cup problem involves a non-replacement situation with reference to which the only conceivable meaning one can give to his infinite population is precisely the postulate he repudiates in the passage cited above, viz. an infinite succession of 4-fold trials without replacement from 2-class universes of 8 objects.

The index of the 1948 edition of the *Statistical Methods* does not cite any page reference against *significance*, *null* or *hypothesis*, but gives p. 41 under *Tests, Significance of*. With this clue and from elsewhere in the same source we may try to get a clear idea of a theory of decision test procedure and/or estimation alternative to that of Neyman and Pearson. I shall first cite the only relevant (p. 41) indication the index supplies :

(i) The idea of an infinite population distributed in a *frequency distribution* in respect of one or more characters is fundamental to all statistical work. From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the infinite hypothetical population from which our sample is drawn, and so of the probable nature of future samples to which our conclusions are to be applied. If a second sample belies this expectation we infer that it is, in the language of statistics, drawn from a different population; that the treatment to which the second sample of organisms had been exposed did in fact make a material difference, or that the climate (or the methods of measuring it) had materially altered. Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first.

Since we have here no definitive criterion of when a sample score belies our expectation, we must seek further afield for an elucidation of what a decision test can accomplish. Three pages later we come to the next unequivocally relevant statement which follows a definition of a normal score deviation in any infinitesimal range  $dx$ .

(ii) In practical applications we do not so often want to know the frequency at any distance from the centre as the total frequency beyond that distance; this is represented by the area of the tail of the curve cut off at any point. Tables of this total frequency, or probability integral, have been constructed from which, for any value of  $(x - \mu)/\sigma$ , we can find what fraction of the total population has a larger deviation; or, in other words, what is the probability that a value so distributed, chosen at random, shall exceed a given deviation. Tables I and II have been constructed to show the deviations corresponding to different values of this probability. The rapidity with which the probability falls off as the deviation increases is well shown in these tables. A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials, while Table II shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which  $P=0.05$ , or 1 in 20 is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

Evidently, it is difficult to confer any meaning on the expressions "about once in 22 trials" or "only once in 370 trials" unless we postulate the extraction as a repetitive process in conformity with the Neyman-Pearson axiom. We might thus conclude that Fisher's objection to the axiom cited arises from the circumstance that it interprets significance levels in terms of frequency of wrong decisions. There is in fact nothing in either of the preceding paragraphs to suggest an alternative view. The

criterion which would lead us "to follow up a false indication only once in 22 trials" presupposes that the sample does in fact come from a particular infinite population; but has no bearing whatsoever on how often we should follow a false indication if the sample came from some other infinite population. The sort of decision to which the test leads us is therefore useful if, and only if, our sole concern is to set a fixed limit to the conditional risk of rejecting the unique proposition implicit in the null hypothesis, viz. that the infinite population from which the sample comes has a certain structure.

The choice of an appropriate null hypothesis then presupposes: (a) that the main preoccupation of the investigator is to safeguard himself or herself against the risk of rejecting as the source of the sample a population with a particular structure; (b) there exists a convenient distribution function definitive of repeated sampling from such a population. If so, the numerous manuals devoted to the exposition of the Fisher battery of tests display a truly astonishing prescience concerning the preoccupations of the research worker or indifference to what they are. Thus a prodigious literature on therapeutic trials records the outcome of a Chi Square *ld.f.* test for the  $2 \times 2$  table in conformity with the null hypothesis that one treatment is as good as another. Actually, the risk the research worker may be most anxious to avoid is that of rejecting a new treatment which is better than the old one, in which event the procedure prescribed by Fisher's pupils and expositors has then no bearing on the presumptive reason for carrying out the test. Clearly, the considerations which dictate the null hypothesis implicit in the test procedures advocated in the many manuals which expound Fisher's methods are mainly referable to what is an essential desideratum of any test procedure, viz. that the sample distribution of the population specified by the hypothesis chosen is definable and amenable to tabulation for ready reckoning. No less evidently, the fact that it may have this desirable property has no necessary connexion with any risk the laboratory or field worker is unwilling to incur.

There is another puzzling feature of the passage first cited, when we interpret it side by side with the second. In the former

Fisher refers to an infinite population as a fundamental postulate of his theory of statistical work. In the latter he elucidates his views about significance levels in terms of a particular population which is *ex hypothesi* infinite; but we could pair off each statement with appropriate verbal changes with a corresponding statement about the risk of following up a false indication about the structure of an urn containing 25 balls some red and some black on the basis of our knowledge that 3 out of 5 balls extracted from it without replacement are red. Surely the reason for introducing the concept of infinity has therefore no special relevance to the population of objects to which the score values of the frequency distribution refer. Its convenience can arise only in a conceptual framework of infinite repetition indispensable to the formulation of a correspondence between assertions and the occurrences falsely or truly described by them.

The rest of the chapter beginning on p. 41 sheds no new light on what Fisher claims for a decision test or on his views about statistical inference in general. On p. 96 we come on the following statement which encourages us to hope for more:

The treatment of frequencies by means of  $\chi^2$  is an approximation, which is useful for the comparative simplicity of the calculations. The exact treatment is somewhat more laborious, though necessary in cases of doubt, and valuable as displaying the true nature of the inferences which the method of  $\chi^2$  is designed to draw.

There follows an exposition of a treatment of  $2 \times 2$  tables to test the hypothesis of proportionality, i.e. that two small samples are samples from populations with the same definitive parameter  $p$ .\* The conclusion is:

Without any assumption or approximation, therefore, the table observed may be judged significantly to contradict the hypothesis of proportionality if

$$\frac{18! \ 13!}{30!} (2992 + 102 + 1)$$

\* Fisher himself does not explicitly in this context make the important semantic distinction between the statement that the two samples come from the same infinite population and the statement that they come from populations which are alike in terms of the classificatory criterion relevant to the methods of sampling.

is a small quantity. This amounts to 619/1330665, or about 1 in 2150 showing that if the hypothesis of proportionality were true, observations of the kind recorded would be highly exceptional.

In this example, the reader will note that Fisher disclaims the necessity to formulate a criterion of rejection before undertaking a decision test; and calculates the frequency with which a particular range of score values, including the sample score itself will turn up, if the null hypothesis is true. No one can take exception to the deduction that "observations of the kind recorded are highly exceptional"; but the logician will wish to know if this conclusion has any necessary bearing on the type of inference relevant to the aim and nature of the test. Kendall's comment on a comparable distribution summarises the most we can say in such a situation:

If this probability is small we have the choice of three possibilities:

- (a) An improbable event has occurred.
- (b) The hypothesis is not true, i.e. the proportion of A's in the population is not  $\bar{w}$ .
- (c) The sampling process is not random.

The calculation cited from p. 97 of *Statistical Methods* refers to "the probabilities of the set of frequencies observed and the two possible more extreme sets of frequencies which might have been observed." If we have so far failed to clarify what role Fisher confers on a significance test as a tool of statistical inference, this *accouplement* at least gets into focus a fundamental difference between his own attitude to the theory of test procedure and that of Neyman, Pearson, Wald, and the Columbia Research Group. Their view is that a decision test embodies a rule with statistically specifiable consequences if followed consistently. The kingpin of the rule is a rejection criterion which is independent of the particular sample under examination. The criterion prescribes a range of sample scores deemed to be inadmissible if the population prescribed by the test hypothesis correctly describes the source of a sample. If the sample score ( $x_0$ ) lies in a *critical region* defined by the vector rejection criterion  $x > x_c$ , any possible value of  $x$  greater than the observed one will also lie in it. If  $x_0$  lies in a critical region

defined by the vector rejection criterion  $x < x_*$ , any possible value of  $x$  less than  $x_0$  will also lie in it.

The prestatement of the rejection criterion thus gives an intelligible meaning to the quadrature which defines the probability ( $\alpha$ ) of rejecting the hypothesis if true; but Fisher offers no alternative justification for the association with  $x_0$  of all values of  $x$  greater than  $x_0$  in the summation cited above, though the exposition of the test procedure is clearly inconsistent with any such prestatement of a uniform rejection criterion. This circumstance seemingly throws more light than does anything elsewhere stated on Fisher's insistent refusal to identify significance levels with frequency of wrong decisions. All he offers in the source he himself cites as a basis for decision is a range of inadmissible score values specified *after* observing the sample on the basis of an intuitive evaluation of its likeability.

*The Design of Experiments*, first published some ten years later than the source cited, gives us an opportunity for exploring the author's second thoughts on significance and on estimation. In the index of the fifth edition (1949) we find the following relevant entries:

*Significance*: 13, 33, 55, 73, 105, 113, 182

*Null Hypothesis*: 15-17, 182-208

*Fiducial Probability*: 182, 195

Of the entries under significance only one explicitly throws further light on the author's viewpoint. In the citation (pp. 13-14) which follows, three different concepts seemingly compete for mastery and somewhat inconclusively.

It is open to the experimenter to be more or less *exacting* in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. Thus, if he wishes to ignore results having probabilities as high as 1 in 20—the probabilities being of course reckoned from the hypothesis that *the phenomenon to be demonstrated is in fact absent*—then it would be useless for him to experiment with only 3 cups of tea of each kind. For 3 objects can be chosen out of 6 in only 20 ways, and there-



fore complete success in the test would be achieved without sensory discrimination, i.e. by "pure chance," in an average of 5 trials out of 100. It is usual and convenient for experiments to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient contention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. *In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Italics inserted.)*

Initial remarks about how exacting the experimenter needs to be re-echo the impression that a significance test as conceived by the author is primarily a disciplinary regimen to discourage rash judgments. The statement that it is *usual* for research workers to adopt a 5 per cent significance level in the same context is true only of those who rely on the many rule of thumb manuals expounding Fisher's own test prescriptions. The explicit admission that we reckon the probabilities on the basis of the truth of *the hypothesis that the phenomenon to be demonstrated is in fact absent* implies that the only type of statement justified by the outcome of the test is conditional on the assumption that the hypothesis is true. In its own context the assertion that we need a *reliable method of procedure* in contradistinction to an isolated record seems to be consistent with the framework of indefinitely protracted repetition postulated by Neyman, von Mises and Wald; but it is the final statement which provokes our attention more especially.

As it stands, its literal form might legitimately convey the

author's intention of introducing a new definition of *demonstrability*; but we do not in fact repeat experiments to see whether they fail to give a significant result in Fisher's sense of the term, or in any other. Accordingly, we must interpret it in its own context *vis-à-vis* the sentence italicised in the preceding paragraph, i.e.: (a) the so-called *null hypothesis* we select as a basis of test procedure is the *negation of the hypothesis which asserts the reality of the phenomenon*; (b) we decline to abandon our suspicion that the alternative hypothesis is true, i.e. that the phenomenon exists, if the null hypothesis assigns an *arbitrarily* low enough probability of occurrence to a specified class of score values including the unique score our single record yields. In any case, we here presume, though in contrariety to the author's apparent intention cited above, a conceptual repetitive framework in which the test procedure operates. Each of the assertions (a) and (b) above thus invite closer scrutiny, to which the author's failure to distinguish a phenomenon from a hypothesis about a phenomenon and his catholic use of the term *variation* adds a special difficulty arising out of the important distinction to which Kendall directs attention, viz: (a) the rarity of the prescribed event in virtue of the hypothetical random distribution implicit in the chosen null hypothesis, *if true*; (b) the rarity of the prescribed event in virtue of departures from randomisation in the assembly of data; (c) the rarity of the prescribed event in virtue of the inapplicability of the null hypothesis to the occurrence.

As regards the author's view of the proper criterion of choice for a unique null hypothesis, several difficulties arise when we ask what form the denial must take, i.e. what precisely do we mean if we assert that the phenomenon to be demonstrated *is in fact absent*? This is indeed the theme (p. 470) of a lengthy and entertaining analysis by Neyman of the lady and the teacup parable referred to for illustrative use in the passage under discussion. The two major difficulties are these: (a) many situations in which the author's expositors could certainly advocate a Fisher test procedure admit of no singular statement of the denial appropriate to the situation; (b) the object of an experiment may be to assess the validity of equally commendable alternatives each being the denial of the other.

A single example will suffice to elucidate (a). We suppose that rival schools, as indeed in the days of the controversy between Weldon and Bateson, assert of a particular class of experiments; (i) a ratio of the most elementary type prescribed by Mendel's hypothesis (3:1) holds good; (ii) such a ratio does not hold good w.r.t. the phenomenon under dispute. Proponents of (ii) may adopt (i) as the denial of their assertion. Proponents of (i) can formulate no such unique denial satisfactory to proponents of (ii). To be sure, they can place themselves in the position of their contestants; but this then imposes the same impracticable obligation on the latter. Few who follow Fisher's test prescriptions appear to realise that there never can be a unique denial of a hypothesis itself formulated, like the theory of the gene, in statistical terms, nor that a 20 to 1 convention can have any relevance to one's assessment of its truth.

In any case, the form of the denial must necessarily satisfy one criterion of adequacy and should rightly satisfy a second. Since a stochastic null hypothesis must prescribe a sample distribution of score values, mathematical tractability of the implicit assumptions in terms of sampling theory will in practice dictate the form that the denial will take. It is therefore of interest to note that the several components of the impressive battery of significance tests associated with Fisher's name consistently imply that the universe of choice is homogeneous w.r.t. the relevant dimension of classification, i.e. that any manifest discrepancy associated with the relevant taxonomical criterion (or criteria) is such as would arise in random sampling from one and the same infinite hypothetical population. Admittedly, this makes the prescription of a test more tractable from the viewpoint of the mathematician; but if there is any other intelligible reason for adopting such a postulate, it is difficult to find an equally intelligible reason for the silence of Fisher's expositors with respect to the operational intention of the denial. Contemporary literature of therapeutic and prophylactic trials is an uninterrupted record of Chi Square tests for  $2 \times 2$  tables to test the null hypothesis that there is no treatment difference, when the question of real interest is the putative existence of a difference sufficient to justify replacement of one treatment by another.

No doubt Fisher would concede that this is an issue of estimation; but this scarcely explains the object in view when performing the so-called exact test cited in his own words above. In practice, the overwhelming majority of biologists and sociologists who employ Fisher's battery of tests have learned the routine as an army drill from the many manuals following the same method of presentation as his *Statistical Methods*. As stated, this gives no derivation of the sampling distributions invoked and no adequate survey of what principles of statistical inference accredit their use. As also stated, the prescribed null hypothesis conforms to a set pattern in which the idea of randomisation is paramount; but Anscombe (1948) has rightly pointed out that this concept admits of interpretation at no less than three semantic levels when conceived in terms of the assertion that *the phenomenon to be demonstrated is not present*.

In the *Design*, the last passage cited is the only one in which the author expounds his fundamental attitude to significance explicitly or at length. As we have seen, it endows the null hypothesis with a unique status which would be less difficult to interpret if the test battery itself supplied the investigator with a range of choice adequate to what denials may be appropriate to the experimental set-up. We therefore turn with bewilderment to what the author has to say (p. 138, *op. cit.*) on this topic in a seemingly oblique reference to the Neyman-Pearson theory of alternative test procedure:

The notion that different tests of significance are appropriate to test different features of the same null hypothesis presents no difficulty to workers engaged in practical experimentation, but has been the occasion of much theoretical discussion among statisticians. The reason for this diversity of viewpoint is perhaps that the experimenter is thinking in terms of observational values, and is aware of what observational discrepancy it is which interests him, and which he thinks may be statistically significant, before he enquires what test of significance, if any, is available appropriate to his needs. He is, therefore, not usually concerned with the question: To what observational feature should a test of significance be applied? This question, when the answer to it is not already known, can be fruitfully discussed only when the experimenter has in view, not a single null hypothesis, but a class of such hypotheses,

in the significance of deviations from each of which he is equally interested. We shall, later, discuss in more detail the logical situation created when this is the case. It should not, however, be thought that such an elaborate theoretical background is a normal condition of experimentation, or that it is needed for the competent and effective use of tests of significance.

Lack of a clear-cut distinction between observational discrepancy and the hypothesis in seeming contrariety to the recorded observation is at this stage a familiar idiom and need not detain us. The passage is quotable since it implies that the laboratory or field worker has no doubts about what is the appropriate null hypothesis to select and has at his or her disposal no lack of appropriate test prescriptions presumably equipped with suitable tables in one of the good books. We have seen that this is not so, and there is no need to add anything to foregoing comments bearing on this charitable interpretation of the consumer's choice and custom. It is not without interest that Fisher's assertion (cited from pp. 13-14) of the unique form the null hypothesis must take is not easy to reconcile with the above and retires from the field on pp. 15-16.

Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations. Tests of significance are of many different kinds, which need not be considered here. Here we are only concerned with the fact that the easy calculation in permutations which we encountered, and which gave us our test of significance, stands for something present in every possible experimental arrangement; or, at least, for something required in its interpretation. The two classes of results which are distinguished by our test of significance are, on the one hand, those which show a significant discrepancy from a certain hypothesis; namely, in this case, the hypothesis that the judgments given are in no way influenced by the order in which the ingredients have been added; and on the other hand, results which show no significant discrepancy from this hypothesis. This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would often be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment we may speak of

this hypothesis as the "null hypothesis," and it should be noted that the *null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation*. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (*Italics inserted.*)

In the light of these remarks and more especially with due regard to the content of the citation from p. 138 of the *Design*, it is permissible to entertain the possibility that Fisher did not initially intend to limit (as on pp. 13-14) the null hypothesis appropriate to every situation by a form of words implying that *the phenomenon to be demonstrated is not present*. Since we cannot explicitly formulate any unique hypothesis conceived as the negation of a particular genetic hypothesis of the Mendelian type, let us therefore explore the consequences of regarding the proper null hypothesis as the Mendelian interpretation itself. Following the prescription of Fisher's school, we should then say that a deviation of  $\pm 2\sigma$  from expectation would be a rare event, but a deviation of  $\pm 3\sigma$  would be more so. Indeed, any deviation numerically greater than  $0.675\sigma$  would be less common than the residual score class. Thus the more exacting we make our significance criterion, the greater will be the chance that a true discrepancy between hypothesis and fact will escape detection. All we can say is: (a) if the hypothesis is true, score deviations within the range  $\pm 0.675\sigma$  will occur as often as score deviations outside this range; (b) if our rejection score criterion  $x_0$  is numerically greater than  $0.675\sigma$  our procedure will more often lead us to correct than to wrong decisions in the long run, *if the hypothesis is true*. Surely it is clear that the conditional risk so specified by the procedure is not necessarily the risk against which the investigator wishes to protect his judgment. Nor is it clear that investigators with opposite views about the *phenomenon to be demonstrated* will assign priority to one and the same alternative risks of recording an erroneous verdict.

This interpretation of an isolated test turns the spotlight on the last words italicised in the citation from pp. 15-16. For reasons set forth in Chapter Thirteen, the assertion that the test cannot demonstrate the truth of the hypothesis, an assertion difficult to dovetail into the intention of the last sentence of

the citation from pp. 13-14, deprives it of any usefulness unless conceived as a disciplinary device; but what precisely entitles us to hope that we have *possibly disproved* it? Within the strait-jacket of the unique null hypothesis, we have merely made a reason for rejecting it dictated by considerations relevant to the penalties of doing so when it happens to be true. In any event, the considerations advanced at the end of the last paragraph suggest that the use of the probable error by a generation with no prescience of the 5 per cent feeling scarcely merits the stricture (*Statistical Methods*, 10th edn., 1948, p. 45).

The value of the deviation beyond which half the observations lie is called the *quartile* distance, and bears to the standard deviation the ratio 0.67449. It was formerly a common practice to calculate the standard error and then, multiplying it by this factor, to obtain the *probable error*. The probable error is thus about two-thirds of the standard error, and as a test of significance a deviation of three times the probable error is effectively equivalent to one of twice the standard error. The common use of the probable error is its only recommendation; when any critical test is required the deviation must be expressed in terms of the standard error in using the tables of normal deviates.

The insertion of italics in this citation is not relevant at this stage. In striking contrast to the pervasive assumption of homogeneity within the framework of test decision is the form of words Fisher uses when he introduces readers of the *Design* (pp. 195-6) to the *fiducial concept of probability that the unknown parameters . . . should be within specified limits*. For the real universe of estimation is in such situations the homogeneous universe of what we elsewhere refer to as Model I; and the probability assignable to a parameter in connexion with anything we say about the interval in which it lies can have only two values, viz. zero and unity.

It is the circumstance that statistics sufficient for the estimation of these two quantities are obtained merely from the sum and the sum of squares of the observations, that gives a peculiar simplicity to problems for which the theory of errors is appropriate. This simplicity appears in an alternative form of statement, which is legitimate in these cases, namely, statements of the *probability that*

*the unknown parameters*, such as  $\mu$  and  $\sigma$ , *should lie within specified limits*. Such statements are termed statements of *fiducial* probability, to distinguish them from the statements of *inverse* probability, by which mathematicians formerly attempted to express the results of inductive inference. Statements of inverse probability have a different logical content from statements of fiducial probability, in spite of their similarity of form, and require for their truth the postulation of knowledge beyond that obtained by direct observation.

In this context it is tempting to cite the curious reasons Fisher gives for restricting the fiducial argument to situations in which we may conceptually invoke a continuous, in contradistinction to a discrete, distribution:

. . . With discontinuous data, however, the fiducial argument only leads to the result that this probability does not exceed 0.01. We have a statement of inequality, and not one of equality. It is not obvious, in such cases, that, of the two forms of statement possible, the one explicitly framed in terms of probability has any practical advantage. The reason why the fiducial statement loses its precision with discontinuous data is that the frequencies in our table make no distinction between a case in which the 2 dizygotic convicts were only just convicted, perhaps on venial charges, or as first offenders, while the remaining 15 had characters above suspicion, and an equally possible case in which the 2 convicts were hardened offenders, and some at least of the remaining 15 had barely escaped conviction. (*The Logic of Inductive Inference*. *J. Roy. Stat. Soc.*, Vol. XCVIII, Pt. I, pp. 50-1, 1935.)

Some of the enigmas and seeming inconsistencies exhibited in previous citations take on a new aspect when we recall: (a) how largely R. A. Fisher by his own admission relies on intuition; (b) how much his later views owe to his early experience, assigned as Statistician to an Agricultural Research Institute with the task of extracting any grain reclaimable from a long-standing accumulation of inexpertly designed field trials. This suffices to explain his preoccupation with *sufficiency*, a concept which has so much less prominence in theories advanced by the opposing school. Insistent concern for what his own school refer to as *Amount of Information* and disdain for excessive consistency go hand in hand in the following from



*The Logic of Inductive Inference* (1935) reprinted in the collected works as 26.47:

. . . One could, therefore, develop a mathematical theory of quantity of information from these properties as postulates, and this would be the normal mathematical procedure. It is, perhaps, only a personal preference that I am more inclined to examine the quantity as it emerges from mathematical investigations, and to judge of its utility by the free use of common sense, rather than to impose it by a formal definition. As a mathematical quantity information is strikingly similar to *entropy* in the mathematical theory of thermodynamics. You will notice especially that reversible processes, changes of notation, mathematical transformations if single-valued, translation of the data into foreign languages, or rewriting them in code, cannot be accompanied by loss of information; but that the irreversible processes involved in statistical estimation, where we cannot reconstruct the original data from the estimate we calculate from it, may be accompanied by a loss, but never by a gain.

The importance of this preoccupation with *amount of information* lies in a basically different orientation of Fisher's sect when we set his views in juxtaposition to those of the alternative school. One mode of thought proposes the question: what rules must we impose on our reasoning before we have permitted the data to influence our views? That of Fisher and his followers asks: what course shall we pursue when we have weighed up all the relevant evidence inherent in the data? Thus we find the following in *Uncertain Inference* (27.254-27.255 *op. cit.*):

. . . There is one peculiarity of uncertain inference which often presents a difficulty to mathematicians trained only in the technique of rigorous deductive argument, namely, that our conclusions are arbitrary, and therefore invalid, unless all the data, exhaustively, are taken into account. In rigorous deductive reasoning we may make any selection from the data, and any certain conclusions which may be deduced from this selection will be valid, whatever additional data we may have at our disposal. . . . This consideration is vital to the fiducial type of argument, which purports to infer exact statements of the probabilities that unknown hypothetical quantities, or that future observations, shall lie within assigned limits, on the basis of a body of observational experience. No such

process could be justified unless the relevant information latent in this experience were exhaustively mobilized and incorporated in our inference.

The special difficulty that arises from Fisher's robust and seemingly contagious confidence in his own intuitions appears in the two following citations, the first of which is from *The Foundations of Theoretical Statistics* (1922) reprinted in the collected works (10.323):

... For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of the maximum likelihood leading in any case to an insufficient statistic. For my own part I should gladly have withheld publication until a rigorously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication, and at the same time I am not insensible of the advantage which accrues to Applied Mathematics from the co-operation of the Pure Mathematician, and this co-operation is not infrequently called forth by the very imperfections of writers on Applied Mathematics.

This intrepid belief in what he disarmingly calls common sense, as a substitute for a system of communicably acceptable rules of procedure, has led Fisher, in a source elsewhere cited, to advance a battery of concepts for the semantic credentials of which neither he nor his disciples offer any justification *en rapport* with generally accepted tenets of the classical theory of probability. Thus an operation which appears in the derivation of Behrens' test as a simple error in terms of the classical theory reappears as a novel and *ad hoc* rule of thought (*vide infra* Yates) described as fiducial inference in Fisher's own treatment of the same issue. Again and again, we seem to sidestep the notion of inverse probability or the invocation of the highly exceptionable Bayes's scholium either by using a new name for the same reasoning process or by ignoring the issue involved.

Thus we come on the following in *Foundations of Theoretical Statistics* reprinted in the collected works (10.326):

There would be no need to emphasize the baseless character of the assumptions made under the titles of inverse probability and BAYES' theorem in view of the decisive criticism to which they have been exposed at the hands of BOOLE, VENN and CHRYSAL, were it not for the fact that the older writers, such as LAPLACE and POISSON, who accepted these assumptions, also laid the foundations of the modern theory of statistics, and have introduced into their discussions of this subject ideas of a similar character. I must indeed plead guilty in my original statement of the Method of the Maximum Likelihood to having based my argument upon the principle of inverse probability; in the same paper, it is true, I emphasized the fact that such inverse probabilities were relative only. That is to say, that while we might speak of one value of  $p$  as having an inverse probability three times that of another value of  $p$ , we might on no account introduce the differential element  $dp$ , so as to be able to say that it was three times as probable that  $p$  should lie in one rather than the other of two equal elements. Upon consideration, therefore, I perceive that the word probability is wrongly used in such a connection; probability is a ratio of frequencies, and about the frequencies of such values we can know nothing whatever. We must return to the actual fact that one value of  $p$ , of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of  $p$ . If we need a word to characterize this relative property of different values of  $p$ , I suggest that we may speak without confusion of the *likelihood* of one value of  $p$  being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity  $p$  would in fact yield the observed sample.

We thus build up a battery of concepts which constitute the exclusive *mystique* of the sect; and a secret language which excludes intercommunication with heretics or schismatics, as in the following from *Inverse Probability* (1930) reprinted in the source already cited (22.532):

... The process of maximizing  $\pi(\phi)$  or  $S(\log \phi)$  is a method of estimation known as the "method of maximum likelihood"; it has in fact no logical connection with inverse probability at all.

The fact that it has been accidentally associated with inverse probability, and that when it is examined objectively in respect of the properties in random sampling of the estimates to which it gives rise, it has shown itself to be of supreme value, are perhaps the sole remaining reasons why that theory is still treated with respect. The function of the  $\theta$ 's maximized is not, however, a probability and does not obey the laws of probability; it involves no differential element  $d\theta_1 d\theta_2 d\theta_3 \dots$ ; it does none the less afford a rational basis for preferring some values of  $\theta$ , or combination of values of the  $\theta$ 's, to others. It is, just as much as a probability, a numerical measure of rational belief, and for that reason is called the *likelihood* of  $\theta_1 \theta_2 \theta_3 \dots$  having given values, to distinguish it from the probability that  $\theta_1 \theta_2 \theta_3 \dots$  lie within assigned limits, since in common speech both terms are loosely used to cover both types of logical situation.

The only notably valiant attempt of one of the apostles of the sect to interpret the teaching of the Leader as a system of logic is that of Yates. In his paper (1939) on *An Apparent Inconsistency arising from Tests of Significance based on Fiducial Distributions of Unknown Parameters* (*Proc. Camb. Phil. Soc.*, Vol. 35) Yates concedes that in Fisher's extension of the Behrens' integral:

We must frankly recognize that we have here introduced a new concept into our methods of inductive inference, which cannot be deduced by the rules of logic from already accepted methods, but which itself requires formal definition.

What follows is less recognisable as a new concept than as a sequence of *ad hoc* assumptions relevant to a particular test prescription. In simple words, the contention is that: (a) the fiducial argument would be invalid if we did not introduce the new concept; (b) the acceptance of the new principle leads to no "inconsistencies" if the estimates are sufficient in Fisher's sense; (c) if we wish to retain the fiducial argument, we may therefore embrace the new concept with a clear conscience. If there is any ulterior reason for doing so, Yates does not explicitly disclose it.

# INDEX

- Additive property, 41, 72  
Agregates, calculus of, 16, 279-316,  
455, 473  
Aitken, 69, 486  
Albert, Prince, of Saxe-Coburg, 19  
Altair (and Procyon), 187, 205  
Amicable, The, 114  
Ancestral inheritance, Law of, 247  
Anscombe, 14, 24, 140, 324, 476, 498  
A priori adequacy, 438, 440, 448  
Aristarchus, 273  
Aristotle, 281  
Averages, locating, 252  
tyranny of, 227  
Avogadro, 283
- Backward Look, the, 328, 370, 377,  
396, 439  
Bacon, 296, 327, 372  
Barnard, G. A., 24  
Bartlett, M. S., 489  
Bateson, 297, 497  
Bayes, Thos., 21, 26, 110, 132, 324,  
435, 504  
Postulate, see Scholium  
Prior Probabilities, 450  
Scholium, 121, 129, 133, 151, 201,  
324, 441, 504  
Theorem, 126, 345  
Begg (and Hogben), 376  
Behrens' Integral, 506  
Test, 443, 504  
Berkeley, 255  
Berkson, J., 438  
Bernard, Claude, 227, 327, 341, 467  
Bernoulli, D., 55, 83, 84, 97, 282  
Bertrand, 17, 172, 181  
Bessel, 187  
Bivariate normal universe, 220  
Boltzmann, 289, 293  
Boole, 139, 505  
Bose, 293  
Bose-Einstein Statistics, 293  
Bowley, 326  
Boyle's Law, 211, 224  
Bradley, 161, 165, 187, 205  
Bridges, 308  
Bronowski, 16  
Brown, Bancroft, 54
- Brunt, 161, 169, 172, 176, 214  
Buffon, 53, 56, 66  
Burbury, 291  
Burnside, 364  
Burt, 271
- Campbell, Norman, 209  
Carnap, 34, 81, 314  
Cattell, R. B., 230  
P-technique, 266  
Causal neutralisation, Gaussian  
Principle of, 62  
Census figures (and Irregular  
Kollektiv), 458  
Central Limit Theorem, 167, 168, 172,  
177, 184, 477-479  
Charles, Enid, 103  
Chevalier de Méré, 36  
Chi Square Test, 359, 491, 497  
Choice, 44  
Chrystal, 61, 505  
Clausius, 283  
Clopper (and Pearson), 444  
Coin tossing, 51, 56, 61, 150, 401, 462  
Columbia Research Group, 493  
Comte, Auguste, 474  
Confidence, Controversy, 400, 433  
Theory of, 399, 439  
Copernicus, 274  
Correns, 297  
Cournot-Westergaard, 179  
Covariance, 481  
Cuenot, 297  
Czuber, 56
- Dahlberg, 307  
D'Alembert, 43, 110  
Darlington, 356  
Darwin, Sir Charles, 103, 175, 256, 286  
David, Miss, 34  
Democritus, 281  
DeMoivre, 111, 159  
DeVries, 297  
Dirac, 293  
Dodge, 25  
Doncaster, Leonard, 235  
Drummond, Henry, 253
- Eddington, 20, 137

- Edgeworth, 178, 220
- Einstein, 205
- Einstein-Bose Statistics, 293
- Eisenhart, Dr. Churchill, 347
- Elderton, 245
- Ellis, Lewis, 60
- Engels, 255
- Epicurus, 281
- Equitable Corporation, 94, 96, 115
- Error Function, 164
  - Gaussian, Theory of, see Gauss.
  - Normal Law of, 177
- Errors, accidental, 215
  - Calculus of, 159
  - Constant, 214
  - Frequency of, 166
  - Meaning of, 214
  - Systematic, 214
- Euler, 95, 111
- Exclusion and Endorsement, Rule of, 405
- Exploration, Calculus of, 17, 210, 456, 467
  
- Factor Analysis, 259
  - Multiple, 261
- Farr, 326
- Feller, 294, 315, 336
- Fermat, 36, 68, 281
- Fermi-Dirac Statistics, 293
- Fick, Evelyn, 247
- Fiducial Probability, see Probability
- Fisher, R. A., 23, 28, 98, 180, 280, 321, 327, 332-342, 441, 458
  - Chi Square Test, 359
  - Fiducial Probability, 399
  - Maximum Likelihood, 504
  - Significance, 487
  - Sufficiency, 206
  - "Tea cup parable", 469, 489
  - Variance, Analysis of, 466
- Fol, 287
- Fourier, 280
- Frequency Approach, 307
  - Distribution, 105
- Fundamental Probability Set, 50
  
- Galton, 18, 103, 174, 178, 179, 210, 233, 242, 247, 298, 308, 325, 467
- Gärtner, 286
- Gassendi, 281, 283
  
- Gauss, 161, 171, 200, 252, 283
  - and Theory of Relativity, 205
  - Causal Neutralisation, 62
  - Error, Theory of, 15, 17, 161, 169-174, 209-223, 237, 241, 321, 328, 364
  - Method of Least Squares, 205
  - Point-estimation, 328, 466
- Gavarret, Jules, 97, 227, 324
- Gene, Theory of the, 383
- Gibbs, Willard, 488
- Gilbert, 273, 284
- Gossart, 20
- Gossett, W. H., 364
- Great Numbers, Law of, 100
- Greenwood, 97, 326, 329, 359
  
- Hagen, 161, 176, 188, 236, 290
- Hagood, M. J., 342
- Haldane, 307
- Halley's Life Table, 84
- Heisenberg, 284
- Heredity, see Mendelism
- Herschel, Sir John, 174
- Hertwig, 287
- Hipparchus, 273
- Hogben, 307
- Hogben (Begg and Hogben), 376
- Hooke, 211, 220, 282
- Hotelling, 269
- Hume, 136
- Huxley, T. H., 258
  
- Induction, Rules of, 430
- Inductive Inference, 146, 503
- Insufficient Reason, Principle of, 35, 139
- Insurance, Life, 94
  - Theory of Probability in, 96
- Interval Estimation, 399
- Irwin, J. O., 24
  
- Jeans, 284, 290
- Jeffreys, Prof., 24, 81, 324, 444
- Jennings, 307
- Jones, Caradog, 18, 326
- Judgments, Calculus of, 21, 281, 314, 319-476
  
- Kelvin, 254
- Kendall, 28, 99, 103, 197, 279, 448, 458, 476, 493, 496
  - and Behrens Test, 444
  - and Irregular Kollektiv, 408

# INDEX

- Kepler, 106, 187  
 Keynes, 18, 19, 54, 57, 63, 102  
 Kinetic Theory of Gases, 16, 227, 279,  
     280, 284, 288, 291  
 King (Wootton and King), 184  
 Kirchhoff, 291  
 Knight, Thos., 286, 299  
 Kollektiv, 75  
     Irregular, 70, 99, 162, 251, 458  
 Kolmogorov, 315  
 Kramer, 336  
 Kuczynski, R. R., 103  
  
 Laird, John, 285  
 Laplace, 15, 21, 44, 58, 92, 97, 130,  
     133-156, 185, 188, 193, 279, 290,  
     324, 441, 468, 505  
 Laxton, 286  
 Lazzerini, 54  
 Least Squares, Method of, 185, 188,  
     189, 193, 197, 205, 213  
 Legendre, 15, 185, 188, 193  
 Leucippus, 281  
 Levy, 35  
 Liapounoff, 159  
 Life Insurance, 94  
 Life Table, Halley's, 84  
     Northampton, 96, 112  
 Linear Regression, 481  
 Lloyd, Prof., 178  
 Lorentz, 205, 290  
 Lotteries, 57  
 Lottery Wheel (Ideal), 458  
 Lottin, Joseph, 19  
 Lucretius, 281  
  
 Malthus, 84, 234, 255  
 Markhof, 200  
 Mathematical Tables, and Irregular  
     Kollektiv, 458  
 Mather, 335  
 Maximum Likelihood, Method of, 197,  
     208, 504  
 Maxwell, Clerk, 16, 279-296  
 Maxwell-Boltzmann Statistics, 293  
 "Mean Man", 178  
 Means, 480  
 Mendel, 235, 279, 297-316, 497, 500  
 Mendelism, 279-316  
 Mendeljeff, 283  
 Meredith, G. P., 273  
  
 Merriman, 194  
 Meyer, 291  
 Mill, J. S., 107, 370, 376, 377  
 Minimum Variance, Principle of, 200  
 Monte-Carlo Roulette, 57  
 Morgan, William, 111, 299, 308  
 Muller, 308  
 Multiplicative Property, 41, 72  
  
 Naudin, 286  
 Needle Problem, 53  
 Newton, 279, 371  
 Neyman, J., 23, 55, 80, 321, 331, 341,  
     366, 371, 399, 415, 420, 433, 441,  
     450, 466, 470  
 Neyman-Pearson Theory of Alternative  
     Test Procedure, 498  
 Northampton Life Table, 96, 112  
 Notation, Determinant, 485  
  
 Occam, William of, 271, 274  
  
 Parameter, The Definitive, 216  
 Pascal, 36, 281  
 Pearl, Raymond, 326  
 Pearson, E. S., 23, 27, 220, 321, 331,  
     341, 366, 399  
 Pearson, Karl, 17, 52, 142, 176, 180,  
     210, 248, 321, 324, 328  
 Peirce, 68  
 Permutations, Master Theorem of, 40  
 Pirogoff, 291  
 Poincaré, 23  
 Point-estimation, 17, 185, 188, 197, 328,  
     356, 399, 466, 473  
     as a Stochastic Procedure, 205  
 Poisson, 505  
 Population, Mendel's Theory of, 16  
 Price, Richard, 111, 131  
 Probability, 44, 63, 64, 96, 435, 454  
     and Human Action, 34  
     Conditional, 117, 125  
     Fiducial, 399, 441  
     Inverse, 21, 111, 299, 505  
     of a Hypothesis, 144  
     of causes, 143  
     Mendelism and, 297-316  
     Posterior, 124, 126, 128, 138, 153  
     Prior, 124, 126, 138, 450  
     Theory of, 138, 307  
     Unconditional, 117

# STATISTICAL THEORY

- Procyon (Altair and), 187, 205  
 Product Moment, 257, 259  
     Correlation Coefficient, 481  
  
 Quetelet, 17-20, 99-109, 136, 161,  
     169-174, 233, 467  
  
 Randomisation, Method of Wrighton,  
     462  
 Regression, 233  
     Coefficient of, 213, 219  
     Theory of, 252, 281  
 Reynolds, 174  
 Ricardo, 255  
 Rietz, 245  
 Romig, 25  
 Roth, 35  
 Roulette, Monte-Carlo, 57  
  
 Scholium, see Bayes  
 Schultz, Henry, 104  
 Smith, 54  
 Smith, Babington, 458  
 Snedecor, 327, 335, 357, 361  
 Spearman, 260, 262, 269, 476  
 Spring, Law of the, 211, 220  
 Stein, 438  
 Stochastic Credibility, 126  
 Stochastic Induction, Patterns of, 421  
 "Student", see Gossett, W. H.  
 Sturtevant, 308  
 Sufficiency, Criterion of, 206  
 Swift, Dean, 118  
 Swinburne, 372  
  
 "Tea cup" Parable, 469-471, 489  
 Thompson, A. T., 458  
 Thomson, Godfrey, 269, 272  
 Thurstone, 257, 268, 274  
  
 Tippett, 458  
 Todhunter, 96  
 Tschermak, 297  
  
 Umpire Bonus Model, 480  
 Uncertainty, Principle of, 284  
 Uncertainty, Safeguard, 126, 345  
 Uspensky, 52, 54, 74, 92  
  
 Van de Waals, 224  
 Vandermonde's Theorem, 41  
 Variances, 480  
 Venn, 61, 63, 70, 325, 329, 505  
 Von Mises, 70, 79, 81, 138, 143, 251,  
     450  
  
 Wald, A., 192, 320, 328, 331, 366, 399,  
     430, 433, 466  
 Wallace, A. R., 326  
 Watt, James, 282  
 Weldon, Prof., 179, 497  
 Whewell, 108  
 Wilks, Prof., 22, 315  
 Williams, N. F. V. M., 230  
 Williamson, 283  
 Wishart, Dr., 27  
 Wolf, 56  
 Wootton (and King), 184  
 Wright, Sir Almroth, 331  
 Wright, Sewell, 307  
 Wrighton, Raymond, 372, 376, 378,  
     399, 433, 451, 452, 462, 470  
     A Priori Adequacy, 438, 440, 448,  
     449  
     Randomising, 462  
  
 Yates, 506  
 Yates (Fisher and), 458  
 Yule, U., 170, 321, 326, 329, 341, 359,  
     476







